

Locally Adaptive Bayes Nonparametric Regression via Nested Gaussian Processes

Bin Zhu and David B. Dunson*

Abstract

We propose a nested Gaussian process (nGP) as a locally adaptive prior for Bayesian nonparametric regression. Specified through a set of stochastic differential equations (SDEs), the nGP imposes a Gaussian process prior for the function's m th-order derivative. The nesting comes in through including a local instantaneous mean function, which is drawn from another Gaussian process inducing adaptivity to locally-varying smoothness. We discuss the support of the nGP prior in terms of the closure of a reproducing kernel Hilbert space, and consider theoretical properties of the posterior. The posterior mean under the nGP prior is shown to be equivalent to the minimizer of a nested penalized sum-of-squares involving penalties for both the global and local roughness of the function. Using highly-efficient Markov chain Monte Carlo for posterior inference, the proposed method performs well in simulation studies compared to several alternatives, and is scalable to massive data, illustrated through a proteomics application.

Key words: Bayesian nonparametric regression; Nested Gaussian processes; Nested smoothing spline; Penalized sum-of-square; Reproducing kernel Hilbert space; Stochastic differential equations.

*Bin Zhu is Postdoctoral Associate, Department of Statistical Science and Center for Human Genetics, Duke University, Durham, NC 27708, (Email: bin.zhu@duke.edu). David B. Dunson is Professor, Department of Statistical Science, Duke University, Durham, NC 27708, (Email: dunson@stat.duke.edu).

1 Introduction

We consider the nonparametric regression problem

$$Y(t) = U(t) + \varepsilon(t), \quad t \in \mathcal{T} = [t_0, t_U], \quad (1)$$

where $U : \mathcal{T} \rightarrow \mathbb{R}$ is an unknown mean regression function to be estimated at $\mathcal{T}_o = \{t_0, t_1, t_2, \dots, t_J < t_U\}$, $t_0 = 0$, and $\boldsymbol{\varepsilon} = [\varepsilon(t_1), \varepsilon(t_2), \dots, \varepsilon(t_J)]' \sim \mathbf{N}_J(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$ a J -dimensional multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\sigma_\varepsilon^2 \mathbf{I}$. We are particularly interested in allowing the smoothness of U to vary locally as a function of t . For example, consider the protein mass spectrometry data in panel (a) of Figure 1. There are clearly regions of t across which the function is very smooth and other regions in which there are distinct spikes, with these spikes being quite important. An additional challenge is that the data are generated in a high-throughput experiment with $J = 11,186$ observations. Hence, we need a statistical model which allows locally-varying smoothness, while also permitting efficient computation even when data are available at a large number of locations along the function.

[Figure 1 about here.]

A commonly used approach for nonparametric regression is to place a Gaussian process (GP) prior (Neal, 1998; Rasmussen and Williams, 2006; Shi and Choi, 2011) on the unknown U , where the GP is usually specified by its mean and covariance function (e.g. squared exponential). The posterior distribution of $U(\mathcal{T}_o)$ can be conveniently obtained as a multivariate Gaussian distribution. When carefully-chosen hyperpriors are placed on the parameters in the covariance kernel, GP priors have been shown to lead to large support, posterior consistency (Ghosal and Roy, 2006; Choi and Schervish, 2007) and even near minimax optimal adaptive rates of posterior contraction (Van der Vaart and Van Zanten, 2008a). However, the focus of this literature has been on isotropic Gaussian processes, which have a single bandwidth parameter controlling global smoothness, with the contraction rate theory assuming the true function has a single smoothness level. There has been applied work allowing the smoothness of a multivariate regression surface to vary in different directions by using predictor-specific bandwidths in a GP with a squared exponential covariance (Savitsky et al., 2011; Zou et al., 2010). Bhattacharya, Pati, and Dunson (2011) recently showed that a carefully-scaled anisotropic GP leads to minimax optimal adaptive rates in anisotropic function classes including when the true function depends on a subset of the predictors. However, the

focus was on allowing a single smoothness level for each predictor, while our current interest is allowing smoothness to vary locally in nonparametric regression in a single predictor.

There is a rich literature on locally-varying smoothing. One popular approach relies on free knot splines, for which various strategies have been proposed to select the number of knots and their locations, including stepwise forward and/or backward knots selection (Friedman and Silverman, 1989; Friedman, 1991; Luo and Wahba, 1997), accurate knots selection scheme (Zhou and Shen, 2001) and Bayesian knots selection (Smith and Kohn, 1996; Denison et al., 1998; Dimatteo et al., 2001) via Gibbs sampling (George and McCulloch, 1993) or reversible jump Markov chain Monte Carlo (Green, 1995). Although many of these methods perform well in simulations, such free knot approaches tend to be highly computationally demanding making their implementation in massive data sets problematic.

In addition to free knot methods, adaptive penalization approaches have also been proposed. An estimate of U is obtained as the minimizer of a penalized sum of squares including a roughness penalty with a spatially-varying smoothness parameter (Wahba, 1995; Ruppert and Carroll, 2000; Pintore et al., 2006; Crainiceanu et al., 2007). Other smoothness adaptive methods include wavelet shrinkage (Donoho and Johnstone, 1995), local polynomial fitting with variable bandwidth (Fan and Gijbels, 1995), L-spline (Abramovich and Steinberg, 1996; Heckman and Ramsay, 2000), mixture of splines (Wood et al., 2002) and linear combination of kernels with varying bandwidths (Wolpert et al., 2011). The common theme of these approaches is to reduce the constraint on the single smoothness level assumption and to implicitly allow the derivatives of U , a common measurement of the smoothness of U , to vary over t .

In this paper, we instead propose a nested Gaussian process (nGP) prior to explicitly model the expectation of the derivative of U as a function of t and to make full Bayesian inference using an efficient Markov chain Monte Carlo (MCMC) algorithm scalable to massive data. More formally, our nGP prior specifies a GP for U 's m th-order derivative $D^m U$ centered on a local instantaneous mean function $A : \mathcal{T} \rightarrow \mathbb{R}$ which is in turn drawn from another GP. Both GPs are defined by stochastic differential equations (SDEs), related to the method proposed by Zhu et al. (2011). However, Zhu et al. (2011) centered their process on a parametric model, while we instead center on a higher-level GP to allow nonparametric locally-adaptive smoothing. Along with the observation equation (1), SDEs can be reformulated as a state space model (Durbin and Koopman, 2001). This reformulation facilitates the application of simulation smoother (Durbin and Koopman, 2002), an

efficient MCMC algorithm with $\mathcal{O}(J)$ computational complexity which is essential to deal with large scale data. We will show that the nGP prior has large support and its posterior distribution is asymptotically consistent. In addition, the posterior mean or mode of U under the nGP prior can be shown to correspond to the minimizer of a penalized sum of squares with nested penalty functions.

The remainder of the paper is organized as follows. Section 2 defines the nGP prior and discusses some of its properties. Section 3 outlines an efficient Markov chain Monte Carlo (MCMC) algorithm for posterior computation. Section 4 presents simulation studies. The proposed method is applied to a mass spectra dataset in Section 5. Finally, Section 6 contains several concluding remarks and outlines some future directions.

2 Nested Gaussian Process Prior

2.1 Definition and Properties

The nGP defines a GP prior for the mean regression function U and the local instantaneous mean function A through the following SDEs with parameters $\sigma_U \in \mathbb{R}^+$ and $\sigma_A \in \mathbb{R}^+$:

$$D^m U(t) = A(t) + \sigma_U \dot{W}_U(t), \quad m \in \mathbb{N} \geq 2 \quad (2)$$

$$D^n A(t) = \sigma_A \dot{W}_A(t), \quad n \in \mathbb{N} \geq 1 \quad (3)$$

where $\dot{W}_U(t)$ and $\dot{W}_A(t)$ are two independent Gaussian white noise processes with mean function $\mathbf{E}\{\dot{W}_U(t)\} = \mathbf{E}\{\dot{W}_A(t)\} = 0$ and covariance function $\mathbf{E}\{\dot{W}_U(t)\dot{W}_U(t')\} = \mathbf{E}\{\dot{W}_A(t)\dot{W}_A(t')\} = \delta(t - t')$ a delta function. The initial value of U and its derivatives up to order $m - 1$ at $t_0 = 0$ are denoted as $\boldsymbol{\mu} = (\mu_0, \mu_1, \dots, \mu_{m-1})' \sim \mathbf{N}_m(\mathbf{0}, \sigma_\mu^2 \mathbf{I})$. Similarly, the initial values of A and its derivatives till order $n - 1$ at $t_0 = 0$ are denoted as $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_{n-1})' \sim \mathbf{N}_n(\mathbf{0}, \sigma_\alpha^2 \mathbf{I})$. In addition, we assume that $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\dot{W}_U(\cdot)$ and $\dot{W}_A(\cdot)$ are mutually independent. The definition of nGP naturally induces a prior for U with varying smoothness. Indeed, the SDE (2) suggests that $\mathbf{E}\{D^m U(t) \mid A(t)\} = A(t)$. Thus, the smoothness of U , measured by $D^m U$, is expected to be centered on a function A varying over t .

We first recall the definition of the reproducing kernel Hilbert space (RKHS) generated by the zero-mean Gaussian process $W = \{W(t) : t \in \mathcal{T}\}$ and the results on the support of W , which will

be useful to explore the theoretical properties of the nGP prior. Let (Ω, \mathcal{A}, P) be the probability space for W such that for any $t_1, t_2, \dots, t_k \in \mathcal{T}$ with $k \in \mathbb{N}$, $\{W(t_1), W(t_2), \dots, W(t_k)\}'$ follow a zero-mean multivariate normal distribution with covariance matrix induced through the covariance function $\mathcal{K}_W : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R}$, defined by $\mathcal{K}_W(s, t) = E\{W(s)W(t)\}$. The RKHS $\mathcal{H}_{\mathcal{K}_W}$ generated by W is the completion of the linear space of all functions

$$t \mapsto \sum_{i=1}^k a_i \mathcal{K}_W(s_i, t), \quad a_1, \dots, a_k \in \mathbb{R}, s_1, \dots, s_k \in \mathcal{T}, k \in \mathbb{N},$$

with the inner product

$$\left\langle \sum_{i=1}^k a_i \mathcal{K}_W(s_i, \cdot), \sum_{j=1}^l b_j \mathcal{K}_W(t_j, \cdot) \right\rangle_{\mathcal{H}_{\mathcal{K}_W}} = \sum_{i=1}^k \sum_{j=1}^l a_i b_j \mathcal{K}_W(s_i, t_j),$$

which satisfies the reproducing property $f(t) = \langle f, \mathcal{K}_W(t, \cdot) \rangle_{\mathcal{H}_{\mathcal{K}_W}}$ for any $f \in \mathcal{H}_{\mathcal{K}_W} : \mathcal{T} \rightarrow \mathbb{R}$.

With the specification of the RKHS $\mathcal{H}_{\mathcal{K}_W}$, we are able to define the support of W as the closure of $\mathcal{H}_{\mathcal{K}_W}$ (Lemma 5.1, Van der Vaart and Van Zanten, 2008b). We apply this definition to characterize the support of the nGP prior, which is formally stated in Theorem 1. Its proof requires the results of the following lemma.

Lemma 1. *The nested Gaussian process U can be written as $U(t) = \tilde{U}_0(t) + \tilde{U}_1(t) + \tilde{A}_0(t) + \tilde{A}_1(t)$, the summation of mutually independent Gaussian processes with the corresponding mean functions $E\{U_0(t)\} = E\{U_1(t)\} = E\{A_0(t)\} = E\{A_1(t)\} = 0$ and covariance functions*

$$\begin{aligned} \mathcal{K}_{\tilde{U}_0}(s, t) &= \sigma_\mu^2 \mathcal{R}_{\tilde{U}_0}(s, t) = \sigma_\mu^2 \sum_{i=0}^{m-1} \phi_i(s) \phi_i(t), \\ \mathcal{K}_{\tilde{U}_1}(s, t) &= \sigma_U^2 \mathcal{R}_{\tilde{U}_1}(s, t) = \sigma_U^2 \int_{\mathcal{T}} G_m(s, u) G_m(t, u) du, \\ \mathcal{K}_{\tilde{A}_0}(s, t) &= \sigma_\alpha^2 \mathcal{R}_{\tilde{A}_0}(s, t) = \sigma_\alpha^2 \sum_{i=0}^{n-1} \phi_{m+i}(s) \phi_{m+i}(t), \\ \mathcal{K}_{\tilde{A}_1}(s, t) &= \sigma_A^2 \mathcal{R}_{\tilde{A}_1}(s, t) = \sigma_A^2 \int_{\mathcal{T}} G_{m+n}(s, u) G_{m+n}(t, u) du, \end{aligned}$$

respectively, where $\phi_i(t) = \frac{t^i}{i!}$ and $G_m(s, u) = \frac{(s-u)_+^{m-1}}{(m-1)!}$.

The proof is in Appendix A.

Theorem 1. *The support of nested Gaussian process U is the closure of RKHS $\mathcal{H}_{\mathcal{K}_U} = \mathcal{H}_{\mathcal{K}_{\tilde{U}_0}} \oplus \mathcal{H}_{\mathcal{K}_{\tilde{U}_1}} \oplus \mathcal{H}_{\mathcal{K}_{\tilde{A}_0}} \oplus \mathcal{H}_{\mathcal{K}_{\tilde{A}_1}}$, the direct sum of RKHSs $\mathcal{H}_{\mathcal{K}_{\tilde{U}_0}}$, $\mathcal{H}_{\mathcal{K}_{\tilde{U}_1}}$, $\mathcal{H}_{\mathcal{K}_{\tilde{A}_0}}$ and $\mathcal{H}_{\mathcal{K}_{\tilde{A}_1}}$ with reproducing kernels $\mathcal{K}_{\tilde{U}_0}(s, t)$, $\mathcal{K}_{\tilde{U}_1}(s, t)$, $\mathcal{K}_{\tilde{A}_0}(s, t)$ and $\mathcal{K}_{\tilde{A}_1}(s, t)$ respectively.*

The proof is in Appendix A. By Corollary 1, it is of interest to note that $\mathcal{H}_{\mathcal{K}_U}$ includes a subspace $\mathcal{H}_{\mathcal{K}_{\tilde{U}}}$, which is the RKHS for the polynomial smoothing spline (Wahba, 1990, Section 1.5).

Corollary 1. *The support of the Gaussian process $\tilde{U} = \tilde{U}_0 + \tilde{U}_1$ as the prior for polynomial smoothing spline is the closure of RKHS $\mathcal{H}_{\mathcal{K}_{\tilde{U}}} = \mathcal{H}_{\mathcal{K}_{\tilde{U}_0}} \oplus \mathcal{H}_{\mathcal{K}_{\tilde{U}_1}}$ with $\mathcal{H}_{\mathcal{K}_{\tilde{U}}} \subset \mathcal{H}_{\mathcal{K}_U}$.*

The proof is in Appendix A. Hence, it is obvious that the nGP prior includes GP prior for polynomial smoothing spline as a special case when $\sigma_\alpha^2 \rightarrow 0$ and $\sigma_A^2 \rightarrow 0$.

The nGP prior can generate functions U arbitrarily close to any function U_0 in the support of the prior. From Theorem 1 it is clear that the support is large and hence the sample paths from the proposed prior can approximate any function in a broad class. As a stronger property, it is also appealing that the posterior distribution concentrate in arbitrarily small neighborhoods of the true function U_0 which generated the data as the sample size J increases, with this property referred to as posterior consistency. More formally, a prior Π on Θ achieves posterior consistency at the true parameter θ_0 if for any neighborhoods \mathcal{U}_ϵ , the posterior distribution $\Pi(\mathcal{U}_\epsilon \mid Y_1, Y_2, \dots, Y_J) \rightarrow 1$ almost surely under Π_{θ_0} , the true joint distribution of observations $\{Y_j\}_{j=1}^J$. For our case, the parameters $\theta = (U, \sigma_\epsilon)$ lie in the product space $\Theta = \mathcal{H}_{\mathcal{K}_U} \times \mathbb{R}^+$ and have a prior $\Pi_\theta = \Pi_U \times \Pi_{\sigma_\epsilon}$, for which Π_U is an nGP prior for U and Π_{σ_ϵ} is a prior distribution for σ_ϵ . The L_1 neighborhood of $\theta_0 = (U_0, \sigma_{\epsilon,0})$ is defined as $\mathcal{U}_\epsilon = \left\{ (U, \sigma_\epsilon) : \|U - U_0\|_1 = \int_0^{t_U} |U(t) - U_0(t)| dt < \epsilon, |\sigma_\epsilon - \sigma_{\epsilon,0}| < \epsilon \right\}$. We further specify a couple of regularity conditions given by:

Assumption 1. t_j arises according to an infill design: for each $S_j = t_{j+1} - t_j$, there exists a constant $0 < C_d \leq 1$ such that $\max_{1 \leq j < J} S_j < \frac{t_U}{C_d J}$.

Assumption 2. The prior distributions $\Pi_{\sigma_\mu^2}$, $\Pi_{\sigma_U^2}$, $\Pi_{\sigma_\alpha^2}$ and $\Pi_{\sigma_A^2}$ satisfy an exponential tail condition. Specifically, there exist sequences M_J , $\sigma_{\mu,J}^2$, $\sigma_{U,J}^2$, $\sigma_{\alpha,J}^2$ and $\sigma_{A,J}^2$ such that: (i) $\Pi_{\sigma_\mu^2}(\sigma_\mu^2 > \sigma_{\mu,J}^2) = e^{-C_\mu J}$, $\Pi_{\sigma_U^2}(\sigma_U^2 > \sigma_{U,J}^2) = e^{-C_U J}$, $\Pi_{\sigma_\alpha^2}(\sigma_\alpha^2 > \sigma_{\alpha,J}^2) = e^{-C_\alpha J}$ and $\Pi_{\sigma_A^2}(\sigma_A^2 > \sigma_{A,J}^2) = e^{-C_A J}$, for some positive constants C_μ , C_U , C_α and C_A ; (ii) $M_J^2 \sigma_J^{-2} \geq C_g J$, for every $C_g > 0$ and σ_J^{-2} , the minimal element of $\{\sigma_{\mu,J}^{-2}, \sigma_{U,J}^{-2}, \sigma_{\alpha,J}^{-2}, \sigma_{A,J}^{-2}\}$.

Assumption 3. The prior distribution Π_{σ_ϵ} is continuous and the $\sigma_{\epsilon,0}$ lies in the support of Π_{σ_ϵ} .

Under those specifications and regularity conditions, the results on strong posterior consistency for the Bayes nonparametric regression with nGP prior is given as follows.

Theorem 2. Let $\{Y_j\}_{j=1}^J$ be the independent but non-identical observations following normal distributions $\{N_1(U(t_j), \sigma_\varepsilon^2)\}_{j=1}^J$ with unknown mean function U and unknown σ_ε^2 at design points t_1, t_2, \dots, t_J . Suppose U follows an nGP prior and the Assumptions 1, 2 and 3 hold. Then for every $\theta_0 \in \Theta$ and every $\epsilon > 0$,

$$\Pi(\mathcal{U}_\epsilon \mid Y_1, Y_2, \dots, Y_J) \rightarrow 1 \text{ a.s. under } \Pi_{\theta_0}.$$

The proof is based on the strong consistency theorem by Choi and Schervish (2007) and is detailed in Appendix A.

2.2 Connection to Nested Smoothing Spline

We show in Theorem 4 that the posterior mean of U under an nGP prior can be related to the minimizer, namely the nested smoothing spline (nSS) \hat{U} , of the following penalized sum-of-squares with nested penalties,

$$\text{nPSS}(t) = \frac{1}{J} \sum_{j=1}^J \{Y(t_j) - U(t_j)\}^2 + \lambda_U \int_{\mathcal{T}} \{D^m U(t) - A(t)\}^2 dt + \lambda_A \int_{\mathcal{T}} \{D^n A(t)\}^2 dt, \quad (4)$$

where $\lambda_U \in \mathbb{R}^+$ and $\lambda_A \in \mathbb{R}^+$ are the smoothing parameters which control the smoothness of unknown functions $U(t)$ and $A(t)$ respectively. The following Theorem 3 and Corollary 2 provide the explicit forms for nSS, for which the proofs are included in Appendix A.

Theorem 3. The nested smoothing spline $\hat{U}(t)$ has the form

$$\begin{aligned} \hat{U}(t) &= \sum_{i=0}^{m-1} \mu_i \phi_i(t) + \sum_{j=1}^J \nu_j \mathcal{R}_{\tilde{U}_1}(t_j, t) + \sum_{i=0}^{n-1} \alpha_i \phi_{m+i}(t) + \sum_{j=1}^J \beta_j \mathcal{R}_{\tilde{A}_1}(t_j, t) \\ &= \boldsymbol{\mu}' \boldsymbol{\phi}_\mu(t) + \boldsymbol{\nu}' \mathbf{R}_{\tilde{U}}(t) + \boldsymbol{\alpha}' \boldsymbol{\phi}_\alpha(t) + \boldsymbol{\beta}' \mathbf{R}_{\tilde{A}}(t), \end{aligned}$$

where $\boldsymbol{\mu} = (\mu_0, \mu_1, \dots, \mu_{m-1})'$, $\boldsymbol{\nu} = (\nu_1, \nu_1, \dots, \nu_J)'$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_{n-1})'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_J)'$ are the coefficients for the bases

$$\begin{aligned} \boldsymbol{\phi}_\mu(t) &= \{\phi_0(t), \phi_1(t), \dots, \phi_{m-1}(t)\}', \quad \mathbf{R}_{\tilde{U}}(t) = \{\mathcal{R}_{\tilde{U}_1}(t_1, t), \mathcal{R}_{\tilde{U}_1}(t_2, t), \dots, \mathcal{R}_{\tilde{U}_1}(t_J, t)\}', \\ \boldsymbol{\phi}_\alpha(t) &= \{\phi_m(t), \phi_{m+1}(t), \dots, \phi_{m+n-1}(t)\}', \quad \mathbf{R}_{\tilde{A}}(t) = \{\mathcal{R}_{\tilde{A}_1}(t_1, t), \mathcal{R}_{\tilde{A}_1}(t_2, t), \dots, \mathcal{R}_{\tilde{A}_1}(t_J, t)\}'. \end{aligned}$$

In addition, the nested penalized sum-of-squares can be written as

$$\begin{aligned} nPSS(t) = & \frac{1}{J} (\mathbf{Y} - \phi_\mu \boldsymbol{\mu} - \mathbf{R}_{\tilde{U}} \boldsymbol{\nu} - \phi_\alpha \boldsymbol{\alpha} - \mathbf{R}_{\tilde{A}} \boldsymbol{\beta})' (\mathbf{Y} - \phi_\mu \boldsymbol{\mu} - \mathbf{R}_{\tilde{U}} \boldsymbol{\nu} - \phi_\alpha \boldsymbol{\alpha} - \mathbf{R}_{\tilde{A}} \boldsymbol{\beta}) \\ & + \lambda_U \boldsymbol{\nu}' \mathbf{R}_{\tilde{U}} \boldsymbol{\nu} + \lambda_A \boldsymbol{\beta}' \mathbf{R}_{\tilde{A}} \boldsymbol{\beta}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{Y} &= \{Y(t_1), Y(t_1), \dots, Y(t_J)\}', \\ \phi_\mu &= \{\phi_\mu(t_1), \phi_\mu(t_2), \dots, \phi_\mu(t_J)\}', \quad \phi_\alpha = \{\phi_\alpha(t_1), \phi_\alpha(t_2), \dots, \phi_\alpha(t_J)\}', \\ \mathbf{R}_{\tilde{U}} &= \{\mathbf{R}_{\tilde{U}}(t_1), \mathbf{R}_{\tilde{U}}(t_2), \dots, \mathbf{R}_{\tilde{U}}(t_J)\}, \quad \mathbf{R}_{\tilde{A}} = \{\mathbf{R}_{\tilde{A}}(t_1), \mathbf{R}_{\tilde{A}}(t_2), \dots, \mathbf{R}_{\tilde{A}}(t_J)\}. \end{aligned}$$

Corollary 2. The coefficients $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of the nested smoothing spline $\hat{U}(t)$ in Theorem 3 are given as

$$\begin{aligned} \boldsymbol{\mu} &= \Sigma_{\mu|\alpha}^{-1} \phi_{\mu|\alpha} \mathbf{S}^{-1} \mathbf{Y}, \\ \boldsymbol{\nu} &= \mathbf{S}^{-1} \left\{ \mathbf{I} - \left(\phi_\mu \Sigma_{\mu|\alpha}^{-1} \phi_{\mu|\alpha} + \phi_\alpha \Sigma_{\alpha|\mu}^{-1} \phi_{\alpha|\mu} \right) \mathbf{S}^{-1} \right\} \mathbf{Y}, \\ \boldsymbol{\alpha} &= \Sigma_{\alpha|\mu}^{-1} \phi_{\alpha|\mu} \mathbf{S}^{-1} \mathbf{Y}, \\ \boldsymbol{\beta} &= \frac{\lambda_U}{\lambda_A} \boldsymbol{\nu}, \end{aligned}$$

where $\phi_{\mu|\alpha} = \phi'_\mu - \Sigma_{\mu\alpha} \Sigma_{\alpha\alpha}^{-1} \phi'_\alpha$, $\phi_{\alpha|\mu} = \phi'_\alpha - \Sigma_{\alpha\mu} \Sigma_{\mu\mu}^{-1} \phi'_\mu$, $\Sigma_{\mu|\alpha} = \Sigma_{\mu\mu} - \Sigma_{\mu\alpha} \Sigma_{\alpha\alpha}^{-1} \Sigma_{\alpha\mu}$, $\Sigma_{\alpha|\mu} = \Sigma_{\alpha\alpha} - \Sigma_{\alpha\mu} \Sigma_{\mu\mu}^{-1} \Sigma_{\mu\alpha}$, $\Sigma_{\mu\mu} = \phi'_\mu \mathbf{S}^{-1} \phi_\mu$, $\Sigma_{\mu\alpha} = \phi'_\mu \mathbf{S}^{-1} \phi_\alpha$, $\Sigma_{\alpha\mu} = \phi'_\alpha \mathbf{S}^{-1} \phi_\mu$, $\Sigma_{\alpha\alpha} = \phi'_\alpha \mathbf{S}^{-1} \phi_\alpha$ and $\mathbf{S} = \mathbf{M}_{\tilde{U}} + \frac{\lambda_U}{\lambda_A} \mathbf{R}_{\tilde{A}} = \mathbf{R}_{\tilde{U}} + J \lambda_U \mathbf{I} + \frac{\lambda_U}{\lambda_A} \mathbf{R}_{\tilde{A}}$.

Corollary 3. Let $\mathbf{B}_\mu = \Sigma_{\mu|\alpha}^{-1} \phi_{\mu|\alpha} \mathbf{S}^{-1}$, $\mathbf{B}_\nu = \mathbf{S}^{-1} \left\{ \mathbf{I} - \left(\phi_\mu \Sigma_{\mu|\alpha}^{-1} \phi_{\mu|\alpha} + \phi_\alpha \Sigma_{\alpha|\mu}^{-1} \phi_{\alpha|\mu} \right) \mathbf{S}^{-1} \right\}$, $\mathbf{B}_\alpha = \Sigma_{\alpha|\mu}^{-1} \phi_{\alpha|\mu} \mathbf{S}^{-1}$ and $\mathbf{B}_\beta = \frac{\lambda_U}{\lambda_A} \mathbf{B}_\nu$. The nested smoothing spline $\hat{U}(t)$ is a linear smoother, expressed in the matrix form as, $\hat{U} = \mathbf{K}_{\lambda_U, \lambda_A} \mathbf{Y}$, where $\mathbf{K}_{\lambda_U, \lambda_A} = \phi_\mu \mathbf{B}_\mu + \mathbf{R}_{\tilde{U}} \mathbf{B}_\nu + \phi_\alpha \mathbf{B}_\alpha + \mathbf{R}_{\tilde{A}} \mathbf{B}_\beta$.

The proof is straightforward by applying Theorem 3 and Corollary 2. As a linear smoother, nSS estimates the mean function by a linear combination of observations with the weight matrix $\mathbf{K}_{\lambda_U, \lambda_A}$.

Theorem 4 below shows the main result of this subsection, i.e. the posterior mean of \mathbf{U} under the nGP prior is equivalent to the nSS \hat{U} when $\sigma_\mu^2 \rightarrow \infty$ and $\sigma_\alpha^2 \rightarrow \infty$. The proof is in Appendix A and is based on the following results of Lemma 2.

Lemma 2. For the observations $\mathbf{Y} = \{Y(t_1), Y(t_2), \dots, Y(t_J)\}'$ and the nested Gaussian process $U(t)$, we have

$$\begin{aligned} E\{U(t)\} &= 0, \\ E\{\mathbf{Y}\} &= \mathbf{0}, \\ \text{Cov}\{U(t), \mathbf{Y}\} &= \sigma_\mu^2 \phi'_\mu(t) \phi'_\mu + \sigma_U^2 \mathbf{R}'_{\tilde{U}}(t) + \sigma_\alpha^2 \phi'_\alpha(t) \phi'_\alpha + \sigma_A^2 \mathbf{R}'_{\tilde{A}}(t), \\ \text{Cov}\{\mathbf{Y}, \mathbf{Y}\} &= \sigma_\mu^2 \phi_\mu \phi'_\mu + \sigma_U^2 \mathbf{R}_{\tilde{U}} + \sigma_\alpha^2 \phi_\alpha \phi'_\alpha + \sigma_A^2 \mathbf{R}_{\tilde{A}} + \sigma_\varepsilon^2 \mathbf{I}. \end{aligned}$$

Theorem 4. Given observations $\mathbf{Y} = \{Y(t_1), Y(t_2), \dots, Y(t_J)\}'$, the posterior mean of $U(t)$ with nested Gaussian process prior is denoted as $\bar{U}_{\sigma_\mu^2, \sigma_\alpha^2}(t) = E\{U(t) \mid \mathbf{Y}, \sigma_\mu^2, \sigma_\alpha^2, \sigma_\varepsilon^2\}$. We have

$$\lim_{\sigma_\mu^2 \rightarrow \infty} \lim_{\sigma_\alpha^2 \rightarrow \infty} \bar{U}_{\sigma_\mu^2, \sigma_\alpha^2}(t) = \hat{U}(t),$$

where $\hat{U}(t)$ is the nested smoothing spline.

3 Posterior Computation

To complete a Bayesian specification, we choose priors for the initial values, covariance parameters in the nGP and residual variance. In particular, we let $\boldsymbol{\mu} \sim \mathbf{N}_m(\mathbf{0}, \sigma_\mu^2 \mathbf{I})$, $\boldsymbol{\alpha} \sim \mathbf{N}_m(\mathbf{0}, \sigma_\alpha^2 \mathbf{I})$, $\sigma_\varepsilon^2 \sim \text{invGamma}(a, b)$, $\sigma_U^2 \sim \text{invGamma}(a, b)$ and $\sigma_A^2 \sim \text{invGamma}(a, b)$, where $\text{invGamma}(a, b)$ denotes the inverse gamma distribution with shape parameter a and scale parameter b . In the applications shown below, the data are rescaled so that the absolute value of the maximum observation is less than 100. We choose diffuse but proper priors by letting $\sigma_\mu^{-1} = \sigma_\alpha^{-1} = a = b = 0.01$ as a default to allow the data to inform strongly, and have observed good performance in a variety of settings for this choice. In practice, we have found the posterior distributions for these hyperparameters to be substantially more concentrated than the prior in applications we have considered, suggesting substantial Bayesian learning.

With this prior specification, we propose an MCMC algorithm for posterior computation. This algorithm consists of two iterative steps: (1) Given the σ_ε^2 , σ_U^2 , σ_A^2 and \mathbf{Y} , draw posterior samples of $\boldsymbol{\mu}$, $\mathbf{U} = \{U(t_1), U(t_2), \dots, U(t_J)\}'$, $\boldsymbol{\alpha}$ and $\mathbf{A} = \{A(t_1), A(t_2), \dots, A(t_J)\}'$; (2) Given the $\boldsymbol{\mu}$, \mathbf{U} , $\boldsymbol{\alpha}$, \mathbf{A} and \mathbf{Y} , draw posterior samples of σ_ε^2 , σ_U^2 and σ_A^2 .

In the first step, it would seem natural to draw \mathbf{U} and \mathbf{A} from their multivariate normal conditional posterior distributions. However, this is extremely expensive computationally in high di-

mensions involving $O(J^3)$ computations in inverting $J \times J$ covariance matrices, which do not have any sparsity structure that can be exploited. To reduce this computational bottleneck in GP models, there is a rich literature relying on low rank matrix approximations (Smola and Bartlett, 2001; Lawrence et al., 2002; Quinonero-Candela and Rasmussen, 2005). Of course, such low rank approximations introduce some associated approximation error, with the magnitude of this error unknown but potentially substantial in our motivating mass spectrometry applications, as it is not clear that typical approximations having sufficiently low rank to be computationally feasible can be accurate.

To bypass the need for such approximations, we propose a different approach that does not require inverting $J \times J$ covariance matrices but instead exploits the Markovian property implied by SDEs (2) and (3). The Markovian property is represented by a stochastic difference equation, namely the state equation, as illustrated for the case when $m = 2$ and $n = 1$ in Proposition 1 which is easily extended to cases with higher order of m and n .

Proposition 1. *When $m = 2$ and $n = 1$, nested Gaussian process $U(t)$ along with its first order derivative $D^1U(t)$ and $A(t)$ follow the state equation:*

$$\boldsymbol{\theta}_{j+1} = \mathbf{G}_j \boldsymbol{\theta}_j + \boldsymbol{\omega}_j,$$

$$\text{where } \boldsymbol{\theta}_{j+1} = \{U(t_{j+1}), D^1U(t_{j+1}), A(t_{j+1})\}', \boldsymbol{\omega}_j \sim N_3(\mathbf{0}, \mathbf{W}_j), \mathbf{G}_j = \begin{pmatrix} 1 & \delta_j & \frac{\delta_j^2}{2} \\ 0 & 1 & \delta_j \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \mathbf{W}_j = \begin{pmatrix} \frac{\delta_j^3}{3}\sigma_U^2 + \frac{\delta_j^5}{20}\sigma_A^2 & \frac{\delta_j^2}{2}\sigma_U^2 + \frac{\delta_j^4}{8}\sigma_A^2 & \frac{\delta_j^3}{6}\sigma_A^2 \\ \frac{\delta_j^2}{2}\sigma_U^2 + \frac{\delta_j^4}{8}\sigma_A^2 & \delta_j\sigma_U^2 + \frac{\delta_j^3}{3}\sigma_A^2 & \frac{\delta_j^2}{2}\sigma_A^2 \\ \frac{\delta_j^3}{6}\sigma_A^2 & \frac{\delta_j^2}{2}\sigma_A^2 & \delta_j\sigma_A^2 \end{pmatrix} \text{ with } \delta_j = t_{j+1} - t_j.$$

The proof is in Appendix A. The state equation combined with the observation equation (1) forms a state space model (West and Harrison, 1997; Durbin and Koopman, 2001), for which the latent states $\boldsymbol{\theta}_j$'s can be efficiently sampled by a simulation smoother algorithm (Durbin and Koopman, 2002) with $O(J)$ computation complexity.

Given the $\boldsymbol{\mu}$, \mathbf{U} , $\boldsymbol{\alpha}$ and \mathbf{A} , posterior samples of σ_ϵ^2 can be obtained by drawing from the inverse-gamma conditional posterior while σ_U^2 and σ_A^2 can be updated in Metropolis-Hastings (MH) steps. We have found that typical MH random walk steps tend to be sticky and it is preferable to use MH independence chain proposals in which one samples candidates for σ_U^2 and σ_A^2 from approximations

to their conditional posteriors that are easy to sample from. To accomplish this, we rely on the following proposition.

Proposition 2. *When δ_j is sufficient small, the state equation in Proposition 1 can be approximated by*

$$\boldsymbol{\theta}_{j+1} = \tilde{\mathbf{G}}_j \boldsymbol{\theta}_j + \tilde{\mathbf{H}}_j \tilde{\boldsymbol{\omega}}_j,$$

where $\tilde{\boldsymbol{\omega}}_j \sim N_2(\mathbf{0}, \tilde{\mathbf{W}}_j)$, $\tilde{\mathbf{G}}_j = \begin{pmatrix} 1 & \delta_j & 0 \\ 0 & 1 & \delta_j \\ 0 & 0 & 1 \end{pmatrix}$, $\tilde{\mathbf{H}}_j = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\tilde{\mathbf{W}}_j = \begin{pmatrix} \sigma_U^2 \delta_j & 0 \\ 0 & \sigma_A^2 \delta_j \end{pmatrix}$.

The above approximate state equation is derived by applying the Euler approximation (chapter 9, Kloeden and Platen, 1992), essentially a first-order Taylor approximation, to the SDEs (2) and (3). Given the $\boldsymbol{\theta}_j$'s, the σ_U^2 and σ_A^2 in the above approximate state equation can be easily sampled as a Bayesian linear regression model with the given coefficients.

Finally, we outline the proposed MCMC algorithm as follows:

- (1). For the state space model with the observation equation (1) and the state equation in Proposition 1, update the latent states $\boldsymbol{\mu}$, \mathbf{U} , $\boldsymbol{\alpha}$ and \mathbf{A} by using the simulation smoother.
- (2). Sample σ_ε^2 from the posterior distribution $\text{invGamma}\left(a + \frac{1}{2}J, b + \frac{1}{2} \sum_{j=1}^J \{Y(t_j) - U(t_j)\}^2\right)$.
- (3a). Given σ_ε^2 , σ_U^2 and σ_A^2 , we sample the latent states $\boldsymbol{\mu}^*$, \mathbf{U}^* , $\boldsymbol{\alpha}^*$ and \mathbf{A}^* for the approximate state space model with the observation equation (1) and the approximate state equation specified in Proposition 2.
- (3b). Given $\boldsymbol{\mu}^*$, \mathbf{U}^* , $\boldsymbol{\alpha}^*$ and \mathbf{A}^* , the proposal σ_U^{2*} and σ_A^{2*} is drawn from the posterior distributions $\text{invGamma}\left(a + \frac{1}{2}J, b + \frac{1}{2} \sum_{j=0}^{J-1} \frac{\{DU^*(t_{j+1}) - DU^*(t_j) - A^*(t_j)\delta_j\}^2}{\delta_j}\right)$ and $\text{invGamma}\left(a + \frac{1}{2}J, b + \frac{1}{2} \sum_{j=0}^{J-1} \frac{\{A^*(t_{j+1}) - A^*(t_j)\}^2}{\delta_j}\right)$, respectively.
- (3c). The proposal σ_U^{2*} and σ_A^{2*} will be accepted with the probability

$$\min \left\{ \prod_{j=0}^{J-1} \frac{f_{N,3}(\boldsymbol{\theta}_{j+1} - \mathbf{G}_j \boldsymbol{\theta}_j \mid \mathbf{0}, \mathbf{W}_j^*) f_{N,2}(\tilde{\mathbf{H}}_j(\boldsymbol{\theta}_{j+1}^* - \tilde{\mathbf{G}}_j \boldsymbol{\theta}_j^*) \mid \mathbf{0}, \tilde{\mathbf{W}}_j)}{f_{N,3}(\boldsymbol{\theta}_{j+1} - \mathbf{G}_j \boldsymbol{\theta}_j \mid \mathbf{0}, \mathbf{W}_j) f_{N,2}(\tilde{\mathbf{H}}_j(\boldsymbol{\theta}_{j+1}^* - \tilde{\mathbf{G}}_j \boldsymbol{\theta}_j^*) \mid \mathbf{0}, \tilde{\mathbf{W}}_j^*)}, 1 \right\},$$

where $f_{N,k}(\mathbf{X} \mid \mathbf{0}, \boldsymbol{\Sigma})$ denotes the probability density function of the k -dimensional normal random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$; $\boldsymbol{\theta}_j$, \mathbf{W}_j and $\tilde{\mathbf{W}}_j$ are specified in Proposition 1 and

2; Similar notions hold for θ_j^* , \mathbf{W}_j^* and $\tilde{\mathbf{W}}_j^*$ with μ , U , α , \mathbf{A} , σ_U^2 and σ_A^2 replaced by μ^* , U^* , α^* , \mathbf{A}^* , σ_U^{2*} and σ_A^{2*} correspondingly.

4 Simulations

We conducted a simulation study to assess the performance of the proposed method, Bayesian nonparametric regression via an nGP prior (BNR-nGP), and compared it to several alternative methods: cubic smoothing spline (SS, Wahba, 1990), wavelet method with the soft minimax threshold (Wavelet1, Donoho and Johnstone, 1994), wavelet method with the soft Stein’s unbiased estimate of risk for threshold choice (Wavelet2, Donoho and Johnstone, 1995) and hybrid adaptive splines (HAS, Luo and Wahba, 1997). For BNR-nGP, we take the posterior mean as the estimate, which is based on the draws from the proposed MCMC algorithm with 1,500 iterations, discarding the first 500 as the burn-in stage and saving remaining ones. The other methods are implemented in R (R Development Core Team, 2011), along with the corresponding R packages for Wavelet methods (wmtsa, Constantine and Percival, 2010) and hybrid adaptive splines (bsml, Wu et al., 2011).

Our first simulation study focuses on four functions adapted from Donoho and Johnstone (1994) with different types of locally-varying smoothness. The functions are plotted in Figure 2, for which the smoothness levels vary, for example, abruptly in panel (a) or gradually in panel (d). For each function, equally-spaced observations are obtained with Gaussian noise, for which the signal-to-noise ratio is $\frac{SD(U)}{\sigma_\varepsilon} = 7$. We use the mean squared error (MSE) $\frac{1}{J} \sum_{j=1}^J \{\hat{U}(t_j) - U_0(t_j)\}^2$ to compare the performance of different methods based on 100 replicates. The simulation results are summarized in Table 1. Among all methods, SS performs worst, which is not surprising since it can not adapt to the locally-varying smoothness. Among the remaining methods, BNR-nGP performs well in general for all cases with either the smallest or the second smallest average MSE across 100 replicates. In contrast, Wavelet2 and HAS may perform better for a given function, but their performances are obviously inferior for another function (e.g. Heavisine for Wavelet2 and Doppler for HAS). This suggests the nGP prior is able to adapt to a wide variety of locally-varying smoothness profiles.

We further compare the proposed method and the alternative methods for analyzing mass spectrometry data. The 100 datasets are generated by the ‘virtual mass spectrometer’ (Coombes et al., 2005a), which considers the physical principles of the instrument. One set of these simulated data is

plotted in Figure 3 with $\sigma_\varepsilon = 66$. The simulated data have been shown to accurately mimic real data (Morris et al., 2005) and are available at <http://bioinformatics.mdanderson.org/Supplements/Datasets/>. Since the analysis of all observations ($J=20,695$) of a given dataset is computational infeasible for HAS, we focus on the analysis of the observations within two regions with $5 < km/z < 8$ (region 1 with $J=2,524$) and $20 < km/z < 25$ (region 2 with $J=2,235$) respectively. Those two regions represent the unique feature of mass spectrometry data. More specifically, with smaller km/z values the peaks are much taller and sharper than the peaks in the region with larger km/z values. The results in Table 1 indicate that the BNR-nGP performs better than the other smoothness adaptive methods for both regions in terms of smaller average MSE and narrower interquartile range of MSE. Although the smoothing spline seems to work well with smaller average MSE in region 2, the peaks are clearly over-smoothed, leading to large MSEs at these important locations. In contrast, BNR-nGP had excellent performance relative to the competitors across locations.

[Figure 2 about here.]

[Figure 3 about here.]

[Table 1 about here.]

5 Applications

We apply the proposed method to protein mass spectrometry (MS) data. Protein MS plays an important role in proteomics for identifying disease-related proteins in the samples (Cottrell and London, 1999; Tibshirani et al., 2004; Domon and Aebersold, 2006; Morris et al., 2008). For example, Panel (a) of Figure 1 plots 11,186 intensities in a pooled sample of nipple aspirate fluid from healthy breasts and breasts with cancer versus the mass to charge ratio m/z of ions (Coombes et al., 2005b). Analysis of protein MS data involves several steps, including spectra alignment, signal extraction, baseline subtraction, normalization and peak detection. As an illustration of our method, we focus on the second step, i.e., estimate the intensity function adjusted for measurement errors. Peaks in the intensity function may correspond to proteins that differ in the expression levels between cancer and control patients.

We fit the Bayes nonparametric regression with nGP prior and ran the MCMC algorithm for 11,000 iterations with the first 1000 iterations discarded as burn-in and every 10th draw retained

for analysis. The trace plots and autocorrelation plots suggested the algorithm converged fast and mixed well. Panel (b) of Figure 1 plots the posterior mean of U and its pointwise 95% credible interval. Note that the posterior mean of U is adapted to the various smoothness at different regions, which is more apparently illustrated by the Panel (c) of Figure 1. Panel (d) of Figure 1 demonstrates the posterior mean and 95% credible interval of rate of intensity change DU , which suggests a peak around 4 km/z.

6 Discussion

We have proposed a novel nested Gaussian process prior, which is designed for flexible nonparametric locally adaptive smoothing while facilitating efficient computation even in large data sets. Most approaches for Bayesian locally adaptive smoothing, such as free knot splines and kernel regression with varying bandwidths, encounter substantial problems with scalability. Even isotropic Gaussian processes, which provide a widely used and studied prior for nonparametric regression, face well known issues in large data sets, with standard approaches for speeding up computation relying on low rank approximations. It is typically not possible to assess the accuracy of such approximations and whether a low rank assumption is warranted for a particular data set. However, when the function of interest is not smooth but can have very many local bumps and features, high resolution data may be intrinsically needed to obtain an accurate estimate of local features of the function, with low rank approximations having poor accuracy. This seems to be the case in mass spectroscopy applications, such as the motivating proteomics example we considered in Section 5. We have simultaneously addressed two fundamental limitations of typical isotropic Gaussian process priors for nonparametric Bayes regression: (i) the lack of spatially-varying smoothness; and (ii) the lack of scalability to large sample sizes. In addition, this was accomplished in a single coherent Bayesian probability model that fully accounts for uncertainty in the function without relying on multistage estimation.

Although we have provided an initial study of some basic theoretical properties, the fundamental motivation in this paper is to obtain a practically useful method. We hope that this initial work stimulates additional research along several interesting lines. The first relates to generalizing the models and computational algorithms to multivariate regression surfaces. Seemingly this will be straightforward to accomplish using additive models and tensor product specifications. The second

is to allow for the incorporation of prior knowledge regarding the shapes of the functions; in some applications, there is information available in the form of differential equations or even a rough knowledge of the types of curves one anticipates, which could ideally be incorporated into an nGP prior. Finally, there are several interesting theoretical directions, such as showing rates of posterior contraction for true functions belonging to a spatially-varying smoothness class.

A Appendix: Proofs of Theoretical Results

A.1 Proof of Lemma 1

We specify $U(t) = \tilde{U}(t) + \tilde{A}(t)$ and $A(t) = D^m \tilde{A}(t)$. By SDEs (2) and (3),

$$D^m \tilde{U}(t) = \sigma_U \dot{W}_U(t), \quad (5)$$

$$D^{m+n} \tilde{A}(t) = \sigma_A \dot{W}_A(t). \quad (6)$$

By applying stochastic integration to SDEs (5) and (6), it can be shown that

$$\begin{aligned} \tilde{U}(t) &= \tilde{U}_0(t) + \tilde{U}_1(t) = \sum_{i=0}^{m-1} \mu_i \phi_i(t) + \sigma_U^2 \int_{\mathcal{T}} G_m(t, u) \dot{W}_U(u) du, \\ \tilde{A}(t) &= \tilde{A}_0(t) + \tilde{A}_1(t) = \sum_{i=0}^{n-1} \alpha_i \phi_{m+i}(t) + \sigma_A^2 \int_{\mathcal{T}} G_{m+n}(t, u) \dot{W}_A(u) du \end{aligned}$$

given the initial values $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$. Since $\tilde{U}_0(t)$, $\tilde{U}_1(t)$, $\tilde{A}_0(t)$ and $\tilde{A}_1(t)$ are the linear combination of Gaussian random variables at every t , they are Gaussian processes defined over t , whose mean functions and covariance functions can be easily derived as required. In addition, $\tilde{U}_0(t)$, $\tilde{U}_1(t)$, $\tilde{A}_0(t)$ and $\tilde{A}_1(t)$ are mutually independent due to the mutually independent assumption of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\dot{W}_U(\cdot)$ and $\dot{W}_A(\cdot)$ in the definition of nGP.

A.2 Proof of Theorem 1

We aim to characterize $\mathcal{H}_{\mathcal{K}_U}$, the RKHS of U with the reproducing kernel $\mathcal{K}_U(s, t)$. The support of U , a mean-zero Gaussian random element, is the closure of $\mathcal{H}_{\mathcal{K}_U}$ (Van der Vaart and Van Zanten, 2008b, Lemma 5.1).

By Loève's Theorem (Berlinet and Thomas-Agnan, 2004, Theorem 35), the RKHSs generated by the processes $\tilde{U}_0(t)$, $\tilde{U}_1(t)$, $\tilde{A}_0(t)$ and $\tilde{A}_1(t)$ with covariance functions $\mathcal{K}_{\tilde{U}_0}(s, t)$, $\mathcal{K}_{\tilde{U}_1}(s, t)$, $\mathcal{K}_{\tilde{A}_0}(s, t)$

and $\mathcal{K}_{\tilde{A}_1}(s, t)$ (given in Lemma 1) are congruent to RKHSs $\mathcal{H}_{\mathcal{K}_{\tilde{U}_0}}$, $\mathcal{H}_{\mathcal{K}_{\tilde{U}_1}}$, $\mathcal{H}_{\mathcal{K}_{\tilde{A}_0}}$ and $\mathcal{H}_{\mathcal{K}_{\tilde{A}_1}}$, respectively. Based on Theorem 5 of Berline and Thomas-Agnan (2004), we conclude that $\mathcal{K}_U(s, t) = \mathcal{K}_{\tilde{U}_0}(s, t) + \mathcal{K}_{\tilde{U}_1}(s, t) + \mathcal{K}_{\tilde{A}_0}(s, t) + \mathcal{K}_{\tilde{A}_1}(s, t)$ is the reproducing kernel of the RKHS

$$\begin{aligned}\mathcal{H}_{\mathcal{K}_U} &= \mathcal{H}_{\mathcal{K}_{\tilde{U}_0}} \oplus \mathcal{H}_{\mathcal{K}_{\tilde{U}_1}} \oplus \mathcal{H}_{\mathcal{K}_{\tilde{A}_0}} \oplus \mathcal{H}_{\mathcal{K}_{\tilde{A}_1}} \\ &= \{U(t) : U(t) = \tilde{U}_0(t) + \tilde{U}_1(t) + \tilde{A}_0(t) + \tilde{A}_1(t), \\ &\quad \tilde{U}_0(t) \in \mathcal{H}_{\mathcal{K}_{\tilde{U}_0}}, \tilde{U}_1(t) \in \mathcal{H}_{\mathcal{K}_{\tilde{U}_1}}, \tilde{A}_0(t) \in \mathcal{H}_{\mathcal{K}_{\tilde{A}_0}}, \tilde{A}_1(t) \in \mathcal{H}_{\mathcal{K}_{\tilde{A}_1}}\}\end{aligned}$$

A.3 Proof of Theorem 2

Similar to the proof of Lemma 1, we specify $U = \tilde{U}_0 + \tilde{U}_1 + \tilde{A}_0 + \tilde{A}_1$, which is a mean zero Gaussian process with continuous and differentiable covariance function $\mathcal{K}_U(s, t) = \mathcal{K}_{\tilde{U}_0}(s, t) + \mathcal{K}_{\tilde{U}_1}(s, t) + \mathcal{K}_{\tilde{A}_0}(s, t) + \mathcal{K}_{\tilde{A}_1}(s, t)$.

We aim to verify the sufficient conditions of the strong consistency theorem (Theorem 1, Choi and Schervish, 2007) for nonparametric regression: (I) prior positivity of neighborhoods and (II) existence of uniformly exponentially consistent tests and sieves Θ_J with $\Pi_U(\Theta_J^C) \leq C_1 \exp(-C_2 J)$ for some positive constants C_1 and C_2 .

Given U is a Gaussian process with continuous sample path and continuous covariance function, it follows from Theorem 4 of Ghosal and Roy (2006) that $\Pi_U(\|U - U_0\|_\infty < \delta) > 0$ for any $\delta > 0$. In addition, for every $\delta > 0$, $\Pi_{\sigma_\varepsilon}\left(\left|\frac{\sigma_\varepsilon}{\sigma_{\varepsilon,0}} - 1\right| < \delta\right) > 0$ under Assumption 3. Hence, we can define a neighborhood $B_\delta = \left\{(U, \sigma_\varepsilon) : \|U - U_0\|_\infty < \delta, \left|\frac{\sigma_\varepsilon}{\sigma_{\varepsilon,0}} - 1\right| < \delta\right\}$ such that $\Pi_{(U, \sigma_\varepsilon)}(B_\delta) > 0$ satisfying the condition (I).

From Theorem 2 of Choi and Schervish (2007), we can show that for a sequence of M_J , there exist uniformly exponentially consistent tests for the sieves $\Theta_J = \{U : \|U\|_\infty < M_J, \|DU\|_\infty < M_J\}$ under the infill design Assumption 1. What remains is to verify the exponentially small probability of $\Theta_{J,0}^C = \{U : \|U\|_\infty > M_J\}$ and $\Theta_{J,1}^C = \{U : \|DU\|_\infty > M_J\}$. Using Borell's inequality (Proposition A.2.7, Van der Vaart and Wellner, 1996), we have

$$\Pi_U(\|U\|_\infty > M_J) \leq C_1 \exp\left(-\frac{C_3 M_J^2}{\sigma^2}\right)$$

for some positive constants C_1 and C_3 , and

$$\begin{aligned}
\sigma^2 &:= \sup_{t \in [0, t_U]} \left\{ E(\tilde{U}_0 + \tilde{U}_1 + \tilde{A}_0 + \tilde{A}_1)^2 \right\} \\
&= \sup_{t \in [0, t_U]} \left\{ E\tilde{U}_0^2 + E\tilde{U}_1^2 + E\tilde{A}_0^2 + E\tilde{A}_1^2 \right\} \\
&= \sigma_\mu^2 \sum_{i=0}^{m-1} \phi_i^2(t_U) + \frac{\sigma_U^2 t_U^{2m-1}}{(m-1)!(m-1)!(2m-1)} + \\
&\quad \sigma_\alpha^2 \sum_{i=0}^{n-1} \phi_{m+i}^2(t_U) + \frac{\sigma_A^2 t_U^{2m+2n-1}}{(m+n-1)!(m+n-1)!(2m+2n-1)}.
\end{aligned}$$

By applying the Borel-Cantelli theorem, we have

$$\Pi_U(\|U\|_\infty > M_J) \leq C_1 \exp(-C_2 J),$$

almost surely under the exponential tail Assumption 2. By the similar arguments, we can show that $\Pi_U(\|DU\|_\infty > M_J) \leq C_1 \exp(-C_2 J)$.

Hence, the conditions (I) and (II) hold, which leads to the strong consistency for Bayesian nonparametric regression with nGP prior.

A.4 Proof of Corollary 1

Note that $\tilde{U}(t) = \tilde{U}_0(t) + \tilde{U}_1(t) = \sum_{i=0}^{m-1} \mu_i \phi_i(t) + \sigma_U^2 \int_{\mathcal{T}} G_m(t, u) \dot{W}_U(u) du$ is the prior for the polynomial smoothing spline (Wahba, 1990, Section 1.5). By the similar arguments in Theorem 1, we can show that the support of \tilde{U} is the closure of RKHS $\mathcal{H}_{\mathcal{K}_{\tilde{U}}} = \mathcal{H}_{\mathcal{K}_{\tilde{U}_0}} \oplus \mathcal{H}_{\mathcal{K}_{\tilde{U}_1}}$.

Thus, $\mathcal{K}_U(s, t) - \mathcal{K}_{\tilde{U}}(s, t) = \mathcal{K}_{\tilde{A}}(s, t) = \mathcal{H}_{\mathcal{K}_{\tilde{A}_0}} \oplus \mathcal{H}_{\mathcal{K}_{\tilde{A}_1}}$ a nonnegative kernel, which implies that $\mathcal{H}_{\mathcal{K}_{\tilde{U}}} \subset \mathcal{H}_{\mathcal{K}_U}$ by Corollary 4 of Aronszajn (1950).

A.5 Proof of Theorem 3

Let $U(t) = \tilde{U}(t) + \tilde{A}(t)$ and $A(t) = D^m \tilde{A}(t)$. The nested penalized sum-of-square (4) can be written as:

$$\text{nPSS}(t) = \frac{1}{J} \sum_{j=1}^J \left\{ Y(t_j) - \tilde{U}(t_j) - \tilde{A}(t_j) \right\}^2 + \lambda_U \int_{\mathcal{T}} \left\{ D^m \tilde{U}(t) \right\}^2 dt + \lambda_A \int_{\mathcal{T}} \left\{ D^{m+n} \tilde{A}(t) \right\}^2 dt, \quad (7)$$

where $\tilde{U}(t)$ is the m -order polynomial smoothing spline and $\tilde{A}(t)$ is the $(m+n)$ -order polynomial smoothing spline.

By the classical RKHS theory of the polynomial smoothing spline (Wahba, 1990, Section 1.2), there exists a unique decomposition of $\tilde{U}(t)$:

$$\begin{aligned}\tilde{U}(t) &= \tilde{U}_0(t) + \tilde{U}_1(t) \\ &= \sum_{i=0}^{m-1} \mu_i \phi_i(t) + \int_{\mathcal{T}} G_m(t, u) D^m \tilde{U}(t) du\end{aligned}$$

with $\tilde{U}_0(t) \in \mathcal{H}_{\mathcal{R}_{\tilde{U}_0}}$ and $\tilde{U}_1(t) \in \mathcal{H}_{\mathcal{R}_{\tilde{U}_1}}$. $\mathcal{H}_{\mathcal{R}_{\tilde{U}_0}} = \{f(t) : D^m f(t) = 0, t \in \mathcal{T}\}$ nad $\mathcal{H}_{\mathcal{R}_{\tilde{U}_1}} = \{f(t) : D^i f(t) \text{ absolutely continuous for } i = 0, 1, \dots, m-1, D^m f(t) \in \mathcal{L}_2(\mathcal{T})\}$ are the RKHSs with reproducing kernel $\mathcal{R}_{\tilde{U}_0}(s, t)$ and $\mathcal{R}_{\tilde{U}_1}(s, t)$ respectively, where $\phi_i(t)$, $G_m(t, u)$, $\mathcal{R}_{\tilde{U}_0}(s, t)$ and $\mathcal{R}_{\tilde{U}_1}(s, t)$ are defined in Theorem 1 with $\mathcal{L}_2(\mathcal{T}) = \{f(t) : \int_{\mathcal{T}} f^2(t) dt < \infty\}$ the space of squared integrable functions defined on index set \mathcal{T} .

Given $\mathcal{T}_o = \{t_j : j = 1, 2, \dots, J\}$, the $\tilde{U}_1(t) \in \mathcal{H}_{\mathcal{R}_{\tilde{U}_1}}$ can be uniquely written as $\tilde{U}_1(t) = \sum_{j=1}^J \nu_j \mathcal{R}_{\tilde{U}_1}(t_j, t) + \eta_{\tilde{U}_1}(t)$, where $\eta_{\tilde{U}_1}(\cdot) \in \mathcal{H}_{\mathcal{R}_{\tilde{U}_1}}$ orthogonal to $\mathcal{R}_{\tilde{U}_1}(t_j, \cdot)$ with inner product $\langle \mathcal{R}_{\tilde{U}_1}(t_j, \cdot), \eta_{\tilde{U}_1}(\cdot) \rangle_{\mathcal{H}_{\mathcal{R}_{\tilde{U}_1}}} = \int_{\mathcal{T}} D^m \mathcal{R}_{\tilde{U}_1}(t_j, u) D^m \eta_{\tilde{U}_1}(u) du = 0$ for $j = 1, 2, \dots, J$.

As a result,

$$\begin{aligned}\int_{\mathcal{T}} \left\{ D^m \tilde{U}(t) \right\}^2 dt &= \int_{\mathcal{T}} \left[D^m \left\{ \sum_{i=0}^{m-1} \mu_i \phi_i(t) + \sum_{j=1}^J \nu_j \mathcal{R}_{\tilde{U}_1}(t_j, t) + \eta_{\tilde{U}_1}(t) \right\} \right]^2 dt \\ &= \sum_{j=1}^J \sum_{j'=1}^J \nu_j \mathcal{R}_{\tilde{U}_1}(t_j, t_{j'}) \nu_{j'} + \int_{\mathcal{T}} \left\{ D^m \eta_{\tilde{U}_1}(t) \right\}^2 dt \\ &= \boldsymbol{\nu}' \mathbf{R}_{\tilde{U}} \boldsymbol{\nu} + \langle \eta_{\tilde{U}_1}(\cdot), \eta_{\tilde{U}_1}(\cdot) \rangle_{\mathcal{H}_{\mathcal{R}_{\tilde{U}_1}}}.\end{aligned}$$

By similar arguments,

$$\begin{aligned}\tilde{A}(t) &= \tilde{A}_0(t) + \tilde{A}_1(t) \\ &= \sum_{i=0}^{n-1} \alpha_i \phi_{m+i}(t) + \sum_{j=1}^J \beta_j \mathcal{R}_{\tilde{A}_1}(t_j, t) + \eta_{\tilde{A}_1}(t),\end{aligned}$$

and

$$\int_{\mathcal{T}} \left\{ D^m \tilde{A}(t) \right\}^2 dt = \boldsymbol{\beta}' \mathbf{R}_{\tilde{A}} \boldsymbol{\beta} + \langle \eta_{\tilde{A}_1}(\cdot), \eta_{\tilde{A}_1}(\cdot) \rangle_{\mathcal{H}_{\mathcal{R}_{\tilde{A}_1}}},$$

where $\tilde{A}_0(t) \in \mathcal{H}_{\mathcal{R}_{\tilde{A}_0}}$ and $\tilde{A}_1(t) \in \mathcal{H}_{\mathcal{R}_{\tilde{A}_1}}$ with $\mathcal{H}_{\mathcal{R}_{\tilde{A}_0}} = \{f(t) : D^{m+n} f(t) = 0, t \in \mathcal{T}\}$ and $\mathcal{H}_{\mathcal{R}_{\tilde{A}_1}} = \{f(t) : D^i f(t) \text{ absolutely continuous for } i = 0, 1, \dots, m+n-1, D^{m+n} f(t) \in \mathcal{L}_2(\mathcal{T})\}$ the

RKHSs with reproducing kernel $\mathcal{R}_{\tilde{A}_0}(s, t)$ and $\mathcal{R}_{\tilde{A}_1}(s, t)$ respectively; $\eta_{\tilde{A}_1}(\cdot) \in \mathcal{H}_{\mathcal{R}_{\tilde{A}_1}}$ is orthogonal to $\mathcal{R}_{\tilde{A}_1}(t_j, \cdot)$ with inner product $\langle \mathcal{R}_{\tilde{A}_1}(t_j, \cdot), \eta_{\tilde{A}_1}(\cdot) \rangle_{\mathcal{H}_{\mathcal{R}_{\tilde{A}_1}}} = \int_{\mathcal{T}} D^m \mathcal{R}_{\tilde{A}_1}(t_j, u) D^m \eta_{\tilde{A}_1}(u) du = 0$ for $j = 1, 2, \dots, J$.

Note that $\eta_{\tilde{U}_1}(t_j) = \langle \mathcal{R}_{\tilde{U}_1}(t_j, \cdot), \eta_{\tilde{U}_1}(\cdot) \rangle_{\mathcal{H}_{\mathcal{R}_{\tilde{U}_1}}} = 0$ and $\eta_{\tilde{A}_1}(t_j) = \langle \mathcal{R}_{\tilde{A}_1}(t_j, \cdot), \eta_{\tilde{A}_1}(\cdot) \rangle_{\mathcal{H}_{\mathcal{R}_{\tilde{A}_1}}} = 0$ due to the reproducing property of $\mathcal{R}_{\tilde{U}_1}(t_j, \cdot)$ and $\mathcal{R}_{\tilde{A}_1}(t_j, \cdot)$. It then follows from expression (7) that

$$\begin{aligned} \text{nPSS}(t) = & \frac{1}{J} (\mathbf{Y} - \phi_{\mu} \boldsymbol{\mu} - \mathbf{R}_{\tilde{U}} \boldsymbol{\nu} - \phi_{\alpha} \boldsymbol{\alpha} - \mathbf{R}_{\tilde{A}} \boldsymbol{\beta})' (\mathbf{Y} - \phi_{\mu} \boldsymbol{\mu} - \mathbf{R}_{\tilde{U}} \boldsymbol{\nu} - \phi_{\alpha} \boldsymbol{\alpha} - \mathbf{R}_{\tilde{A}} \boldsymbol{\beta}) \\ & + \lambda_U \boldsymbol{\nu}' \mathbf{R}_{\tilde{U}} \boldsymbol{\nu} + \lambda_A \boldsymbol{\beta}' \mathbf{R}_{\tilde{A}} \boldsymbol{\beta} + \langle \eta_{\tilde{U}_1}(\cdot), \eta_{\tilde{U}_1}(\cdot) \rangle_{\mathcal{H}_{\mathcal{R}_{\tilde{U}_1}}} + \langle \eta_{\tilde{A}_1}(\cdot), \eta_{\tilde{A}_1}(\cdot) \rangle_{\mathcal{H}_{\mathcal{R}_{\tilde{A}_1}}}, \end{aligned}$$

which is minimized when $\langle \eta_{\tilde{U}_1}(\cdot), \eta_{\tilde{U}_1}(\cdot) \rangle_{\mathcal{H}_{\mathcal{R}_{\tilde{U}_1}}} = \langle \eta_{\tilde{A}_1}(\cdot), \eta_{\tilde{A}_1}(\cdot) \rangle_{\mathcal{H}_{\mathcal{R}_{\tilde{A}_1}}} = 0$. Thus, $\eta_{\tilde{U}_1}(\cdot) = \eta_{\tilde{A}_1}(\cdot) = 0$ and we obtain the forms of $\hat{U}(t)$ and $\text{nPSS}(t)$ as required.

A.6 Proof of Corollary 2

We first take partial derivatives of nested penalized sum-of-squares $\text{nPSS}(t)$ in Theorem 3 with respect to $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and set them to zeros:

$$\frac{\partial \text{nPSS}(t)}{\partial \boldsymbol{\mu}} = \phi'_{\mu} (\phi_{\mu} \boldsymbol{\mu} + \mathbf{R}_{\tilde{U}} \boldsymbol{\nu} + \phi_{\alpha} \boldsymbol{\alpha} + \mathbf{R}_{\tilde{A}} \boldsymbol{\beta} - \mathbf{Y}) = \mathbf{0}, \quad (8)$$

$$\frac{\partial \text{nPSS}(t)}{\partial \boldsymbol{\nu}} = \mathbf{R}_{\tilde{U}} (\phi_{\mu} \boldsymbol{\mu} + \mathbf{M}_{\tilde{U}} \boldsymbol{\nu} + \phi_{\alpha} \boldsymbol{\alpha} + \mathbf{R}_{\tilde{A}} \boldsymbol{\beta} - \mathbf{Y}) = \mathbf{0}, \quad (9)$$

$$\frac{\partial \text{nPSS}(t)}{\partial \boldsymbol{\alpha}} = \phi'_{\alpha} (\phi_{\mu} \boldsymbol{\mu} + \mathbf{R}_{\tilde{U}} \boldsymbol{\nu} + \phi_{\alpha} \boldsymbol{\alpha} + \mathbf{R}_{\tilde{A}} \boldsymbol{\beta} - \mathbf{Y}) = \mathbf{0}, \quad (10)$$

$$\frac{\partial \text{nPSS}(t)}{\partial \boldsymbol{\beta}} = \mathbf{R}_{\tilde{A}} (\phi_{\mu} \boldsymbol{\mu} + \mathbf{R}_{\tilde{U}} \boldsymbol{\nu} + \phi_{\alpha} \boldsymbol{\alpha} + \mathbf{M}_{\tilde{A}} \boldsymbol{\beta} - \mathbf{Y}) = \mathbf{0}, \quad (11)$$

where $\mathbf{M}_{\tilde{U}} = \mathbf{R}_{\tilde{U}} + J\lambda_U \mathbf{I}$ and $\mathbf{M}_{\tilde{A}} = \mathbf{R}_{\tilde{A}} + J\lambda_A \mathbf{I}$. It follows from equations (9) and (11) that

$$\begin{aligned} \boldsymbol{\nu} &= \mathbf{S}^{-1} (\mathbf{Y} - \phi_{\mu} \boldsymbol{\mu} - \phi_{\alpha} \boldsymbol{\alpha}), \\ \boldsymbol{\beta} &= \frac{\lambda_U}{\lambda_A} \boldsymbol{\nu}. \end{aligned}$$

Substituting them into equations (8) and (10) with some algebra leads to

$$\begin{aligned} \Sigma_{\mu\mu} \boldsymbol{\mu} + \Sigma_{\mu\alpha} \boldsymbol{\alpha} &= \phi'_{\mu} \mathbf{S}^{-1} \mathbf{Y}, \\ \Sigma_{\alpha\mu} \boldsymbol{\mu} + \Sigma_{\alpha\alpha} \boldsymbol{\alpha} &= \phi'_{\alpha} \mathbf{S}^{-1} \mathbf{Y}, \end{aligned}$$

from which we obtain

$$\begin{aligned} \boldsymbol{\mu} &= (\Sigma_{\mu\mu} - \Sigma_{\mu\alpha} \Sigma_{\alpha\alpha}^{-1} \Sigma_{\alpha\mu})^{-1} (\phi'_{\mu} - \Sigma_{\mu\alpha} \Sigma_{\alpha\alpha}^{-1} \phi'_{\alpha}) \mathbf{S}^{-1} \mathbf{Y} = \Sigma_{\mu|\alpha}^{-1} \phi_{\mu|\alpha} \mathbf{S}^{-1} \mathbf{Y}, \\ \boldsymbol{\alpha} &= (\Sigma_{\alpha\alpha} - \Sigma_{\alpha\mu} \Sigma_{\mu\mu}^{-1} \Sigma_{\mu\alpha})^{-1} (\phi'_{\alpha} - \Sigma_{\alpha\mu} \Sigma_{\mu\mu}^{-1} \phi'_{\mu}) \mathbf{S}^{-1} \mathbf{Y} = \Sigma_{\alpha|\mu}^{-1} \phi_{\alpha|\mu} \mathbf{S}^{-1} \mathbf{Y}. \end{aligned}$$

It is then straightforward to show

$$\begin{aligned}\boldsymbol{\nu} &= \mathbf{S}^{-1} \left\{ \mathbf{I} - \left(\boldsymbol{\phi}_\mu \boldsymbol{\Sigma}_{\mu|\alpha}^{-1} \boldsymbol{\phi}_{\mu|\alpha} + \boldsymbol{\phi}_\alpha \boldsymbol{\Sigma}_{\alpha|\mu}^{-1} \boldsymbol{\phi}_{\alpha|\mu} \right) \mathbf{S}^{-1} \right\} \mathbf{Y}, \\ \boldsymbol{\beta} &= \frac{\lambda_U}{\lambda_A} \boldsymbol{\nu}\end{aligned}$$

as desired.

A.7 Proof of Lemma 2

Let $U(t) = \tilde{U}(t) + \tilde{A}(t)$ and $A(t) = D^m \tilde{A}(t)$. From SDEs (2) and (3),

$$D^m \tilde{U}(t) = \sigma_U \dot{W}_U(t),$$

$$D^{m+n} \tilde{A}(t) = \sigma_A \dot{W}_A(t).$$

Thus, given the initial value $\boldsymbol{\mu}$, it can be shown that $\tilde{U}(t) = \sum_{i=0}^{m-1} \mu_i \phi_i(t) + \sigma_U^2 \int_{\mathcal{T}} G_m(t, u) \dot{W}_U(u) du$, a $(m-1)$ -fold integrated Wiener process (Shepp, 1966). Similarly, $\tilde{A}(t) = \sum_{i=0}^{n-1} \alpha_i \phi_{m+i}(t) + \sigma_A^2 \int_{\mathcal{T}} G_{m+n}(t, u) \dot{W}_A(u) du$, a $(m+n-1)$ -fold integrated Wiener process.

It is obvious that $\mathbf{E}\{U(t)\} = 0$ and $\mathbf{E}\{\mathbf{Y}\} = \mathbf{0}$. Given the mutually independent assumption of $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\dot{W}_U(\cdot)$ and $\dot{W}_A(\cdot)$,

$$\begin{aligned}\text{Cov}\{U(t), Y(t_j)\} &= \text{Cov}\{U(t), U(t_j)\} \\ &= \text{Cov}\{\tilde{U}(t), \tilde{U}(t_j)\} + \text{Cov}\{\tilde{A}(t), \tilde{A}(t_j)\} \\ &= \mathbf{E}\{\tilde{U}(t)\tilde{U}(t_j)\} + \mathbf{E}\{\tilde{A}(t)\tilde{A}(t_j)\} \\ &= \sigma_\mu^2 \sum_{i=0}^{m-1} \phi_i(t)\phi_i(t_j) + \sigma_U^2 \mathcal{R}_{\tilde{U}_1}(t, t_j) + \\ &\quad \sigma_\alpha^2 \sum_{i=0}^{n-1} \phi_{m+i}(t)\phi_{m+i}(t_j) + \sigma_A^2 \mathcal{R}_{\tilde{A}_1}(t, t_j),\end{aligned}$$

and

$$\begin{aligned}\text{Cov}\{Y(t_j), Y(t_{j'})\} &= \text{Cov}\{U(t_j), U(t_{j'})\} + \sigma_\varepsilon^2 \\ &= \sigma_\mu^2 \sum_{i=0}^{m-1} \phi_i(t_j)\phi_i(t_{j'}) + \sigma_U^2 \mathcal{R}_{\tilde{U}_1}(t_j, t_{j'}) + \\ &\quad \sigma_\alpha^2 \sum_{i=0}^{n-1} \phi_{m+i}(t_j)\phi_{m+i}(t_{j'}) + \sigma_A^2 \mathcal{R}_{\tilde{A}_1}(t_j, t_{j'}) + \sigma_\varepsilon^2,\end{aligned}$$

for $j = 1, 2, \dots, J$ and $j' = 1, 2, \dots, J$. The lemma holds.

A.8 Proof of Theorem 4

By Lemma 2 and the results on conditional multivariate normal distribution (Searle, 1982),

$$\begin{aligned}
\mathbb{E}\{U(t) \mid \mathbf{Y}, \sigma_\mu^2, \sigma_\alpha^2, \sigma_\varepsilon^2\} &= \text{Cov}\{U(t), \mathbf{Y}\} \text{Cov}^{-1}\{\mathbf{Y}, \mathbf{Y}\} \mathbf{Y} \\
&= [\rho_\mu \phi'_\mu(t) \phi'_\mu + \mathbf{R}'_{\tilde{U}}(t) + \rho_\alpha \phi'_\alpha(t) \phi'_\alpha + \rho_A \mathbf{R}'_{\tilde{A}}(t)] \times \\
&\quad [\rho_\mu \phi_\mu \phi'_\mu + \rho_\alpha \phi_\alpha \phi'_\alpha + \rho_A \mathbf{R}_{\tilde{A}} + \mathbf{R}_{\tilde{U}} + J\lambda_U \mathbf{I}]^{-1} \mathbf{Y} \\
&= \phi'_\mu(t) \left(\rho_\mu \phi'_\mu \Sigma_{\rho_\mu \rho_\alpha}^{-1} \right) \mathbf{Y} + \mathbf{R}'_{\tilde{U}}(t) \Sigma_{\rho_\mu \rho_\alpha}^{-1} \mathbf{Y} + \\
&\quad \phi'_\alpha(t) \left(\rho_\alpha \phi'_\alpha \Sigma_{\rho_\mu \rho_\alpha}^{-1} \right) \mathbf{Y} + \rho_A \mathbf{R}'_{\tilde{A}}(t) \Sigma_{\rho_\mu \rho_\alpha}^{-1} \mathbf{Y}
\end{aligned}$$

where $\rho_\mu = \sigma_\mu^2/\sigma_U^2$, $\rho_\alpha = \sigma_\alpha^2/\sigma_U^2$, $\rho_A = \sigma_A^2/\sigma_U^2$, $J\lambda_U = \sigma_\varepsilon^2/\sigma_U^2$ and $\Sigma_{\rho_\mu \rho_\alpha} = \rho_\mu \phi_\mu \phi'_\mu + \mathbf{S}_{\rho_\alpha}$ with $\mathbf{S}_{\rho_\alpha} = \rho_\alpha \phi_\alpha \phi'_\alpha + \mathbf{S} = \rho_\alpha \phi_\alpha \phi'_\alpha + \rho_A \mathbf{R}_{\tilde{A}} + \mathbf{R}_{\tilde{U}} + J\lambda_U \mathbf{I}$. We are going to evaluate the limits of $\rho_\mu \phi'_\mu \Sigma_{\rho_\mu \rho_\alpha}^{-1}$, $\rho_\alpha \phi'_\alpha \Sigma_{\rho_\mu \rho_\alpha}^{-1}$ and $\Sigma_{\rho_\mu \rho_\alpha}^{-1}$ when $\rho_\mu \rightarrow +\infty$ and $\rho_\alpha \rightarrow +\infty$.

It can be verified that

$$\begin{aligned}
\Sigma_{\rho_\mu \rho_\alpha}^{-1} &= \mathbf{S}_{\rho_\alpha}^{-1} - \mathbf{S}_{\rho_\alpha}^{-1} \phi_\mu (\phi'_\mu \mathbf{S}_{\rho_\alpha}^{-1} \phi_\mu)^{-1} \left\{ \mathbf{I} + \rho_\mu^{-1} (\phi'_\mu \mathbf{S}_{\rho_\alpha}^{-1} \phi_\mu)^{-1} \right\}^{-1} \phi'_\mu \mathbf{S}_{\rho_\alpha}^{-1}, \\
\mathbf{S}_{\rho_\alpha}^{-1} &= \mathbf{S}^{-1} - \mathbf{S}^{-1} \phi_\alpha (\phi'_\alpha \mathbf{S}^{-1} \phi_\alpha)^{-1} \left\{ \mathbf{I} + \rho_\alpha^{-1} (\phi'_\alpha \mathbf{S}^{-1} \phi_\alpha)^{-1} \right\}^{-1} \phi'_\alpha \mathbf{S}^{-1}.
\end{aligned} \tag{12}$$

It follows that $\mathbf{S}_\infty^{-1} = \lim_{\rho_\alpha \rightarrow +\infty} \mathbf{S}_{\rho_\alpha}^{-1} = \mathbf{S}^{-1} - \mathbf{S}^{-1} \phi_\alpha (\phi'_\alpha \mathbf{S}^{-1} \phi_\alpha)^{-1} \phi'_\alpha \mathbf{S}^{-1} = \mathbf{S}^{-1} - \mathbf{S}^{-1} \phi_\alpha \Sigma_{\alpha\alpha}^{-1} \phi'_\alpha \mathbf{S}^{-1}$ and $\phi'_\mu \mathbf{S}_\infty^{-1} \phi_\mu = \Sigma_{\mu|\alpha}$ and $\mathbf{S}_\infty^{-1} \phi_\mu = \mathbf{S}^{-1} \phi'_{\mu|\alpha}$.

As a result,

$$\begin{aligned}
\Sigma_{\infty\infty}^{-1} &= \lim_{\rho_\mu \rightarrow +\infty} \lim_{\rho_\alpha \rightarrow +\infty} \Sigma_{\rho_\mu \rho_\alpha}^{-1} \\
&= \mathbf{S}^{-1} - \mathbf{S}^{-1} \left(\phi_\alpha \Sigma_{\alpha\alpha}^{-1} \phi'_\alpha + \phi'_{\mu|\alpha} \Sigma_{\mu|\alpha}^{-1} \phi_{\mu|\alpha} \right) \mathbf{S}^{-1} \\
&= \mathbf{S}^{-1} - \mathbf{S}^{-1} \left\{ \phi_\mu \Sigma_{\mu|\alpha}^{-1} \phi_{\mu|\alpha} + \phi_\alpha \left(\Sigma_{\alpha\alpha}^{-1} \phi'_\alpha - \Sigma_{\alpha\alpha}^{-1} \Sigma_{\alpha\mu} \Sigma_{\mu|\alpha}^{-1} \phi_{\mu|\alpha} \right) \right\} \mathbf{S}^{-1} \\
&= \mathbf{S}^{-1} \left\{ \mathbf{I} - \left(\phi_\mu \Sigma_{\mu|\alpha}^{-1} \phi_{\mu|\alpha} + \phi_\alpha \Sigma_{\alpha|\mu}^{-1} \phi_{\alpha|\mu} \right) \mathbf{S}^{-1} \right\}.
\end{aligned}$$

By expression (12),

$$\begin{aligned}
\rho_\mu \phi'_\mu \Sigma_{\rho_\mu \rho_\alpha}^{-1} &= \rho_\mu \left[\mathbf{I} - \left\{ \mathbf{I} + \rho_\mu^{-1} (\phi'_\mu \mathbf{S}_{\rho_\alpha}^{-1} \phi_\mu)^{-1} \right\}^{-1} \right] \phi'_\mu \mathbf{S}_{\rho_\alpha}^{-1} \\
&= (\phi'_\mu \mathbf{S}_{\rho_\alpha}^{-1} \phi_\mu)^{-1} \left\{ \mathbf{I} + \rho_\mu^{-1} (\phi'_\mu \mathbf{S}_{\rho_\alpha}^{-1} \phi_\mu)^{-1} \right\}^{-1} \phi'_\mu \mathbf{S}_{\rho_\alpha}^{-1}.
\end{aligned}$$

It follows that $\lim_{\rho_\mu \rightarrow +\infty} \lim_{\rho_\alpha \rightarrow +\infty} \rho_\mu \phi'_\mu \Sigma_{\rho_\mu \rho_\alpha}^{-1} = \Sigma_{\mu|\alpha}^{-1} \phi_{\mu|\alpha} \mathbf{S}^{-1}$. By similar arguments, $\lim_{\rho_\mu \rightarrow +\infty} \lim_{\rho_\alpha \rightarrow +\infty} \rho_\alpha \phi'_\alpha \Sigma_{\rho_\mu \rho_\alpha}^{-1} = \Sigma_{\alpha|\mu}^{-1} \phi_{\alpha|\mu} \mathbf{S}^{-1}$.

Hence, $\Sigma_{\infty\infty}^{-1}\mathbf{Y} = \boldsymbol{\nu}$, $\rho_A \Sigma_{\infty\infty}^{-1}\mathbf{Y} = \boldsymbol{\beta}$, $\lim_{\rho_\mu \rightarrow +\infty} \lim_{\rho_\alpha \rightarrow +\infty} \rho_\mu \phi'_\mu \Sigma_{\rho_\mu \rho_\alpha}^{-1} \mathbf{Y} = \boldsymbol{\mu}$ and $\lim_{\rho_\mu \rightarrow +\infty} \lim_{\rho_\alpha \rightarrow +\infty} \rho_\alpha \phi'_\alpha \Sigma_{\rho_\mu \rho_\alpha}^{-1} \mathbf{Y} = \boldsymbol{\alpha}$. The theorem holds.

A.9 Proof of Proposition 1

When $m = 2$ and $n = 1$, the SDEs (2) and (3) can be written as,

$$D^1 \boldsymbol{\theta}(t) = \mathbf{C} \boldsymbol{\theta}(t) + \mathbf{D} \dot{\mathbf{W}}(t),$$

$$\text{where } \boldsymbol{\theta}(t) = \begin{Bmatrix} U(t) \\ D^1 U(t) \\ A(t) \end{Bmatrix}, \mathbf{C} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \mathbf{D} = \begin{pmatrix} 0 & 0 \\ \sigma_U & 0 \\ 0 & \sigma_A \end{pmatrix} \text{ and } \dot{\mathbf{W}}(t) = \begin{Bmatrix} \dot{W}_U(t) \\ \dot{W}_A(t) \end{Bmatrix}.$$

As a result,

$$\begin{aligned} \boldsymbol{\theta}_{j+1} &= \exp(\mathbf{C} \delta_j) \boldsymbol{\theta}_j + \int_0^{\delta_j} \exp\{\mathbf{C}(\delta_j - u)\} \mathbf{D} \dot{\mathbf{W}}(t_j + u) du \\ &= \mathbf{G}_j \boldsymbol{\theta}_j + \boldsymbol{\omega}_j, \end{aligned}$$

$$\text{where } \mathbf{G}_j = \exp(\mathbf{C} \delta_j) = \mathbf{I} + \delta_j \mathbf{C} + \delta_j^2 \mathbf{C} \mathbf{C} / 2 = \begin{pmatrix} 1 & \delta_j & \frac{\delta_j^2}{2} \\ 0 & 1 & \delta_j \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \boldsymbol{\omega}_j \sim \mathbf{N}_3(\mathbf{0}, \mathbf{W}_j) \text{ with}$$

$$\begin{aligned} \mathbf{W}_j &= \int_0^{\delta_j} \exp\{\mathbf{C}(\delta_j - u)\} \mathbf{D} \mathbf{D}' \exp\{\mathbf{C}'(\delta_j - u)\} du \\ &= \begin{pmatrix} \frac{\delta_j^3}{3} \sigma_U^2 + \frac{\delta_j^5}{20} \sigma_A^2 & \frac{\delta_j^2}{2} \sigma_U^2 + \frac{\delta_j^4}{8} \sigma_A^2 & \frac{\delta_j^3}{6} \sigma_A^2 \\ \frac{\delta_j^2}{2} \sigma_U^2 + \frac{\delta_j^4}{8} \sigma_A^2 & \delta_j \sigma_U^2 + \frac{\delta_j^3}{3} \sigma_A^2 & \frac{\delta_j^2}{2} \sigma_A^2 \\ \frac{\delta_j^3}{6} \sigma_A^2 & \frac{\delta_j^2}{2} \sigma_A^2 & \delta_j \sigma_A^2 \end{pmatrix} \end{aligned}$$

as required.

References

- Abramovich, F. and Steinberg, D.M. (1996), “Improved inference in nonparametric regression using L-smoothing splines,” *Journal of Statistical Planning and Inference*, 49, 327–341.
- Aronszajn, N. (1950), “Theory of Reproducing Kernels,” *Transactions of the American Mathematical Society*, 68, 337–404.

- Berlinet, A. and Thomas-Agnan, C. (2004), *Reproducing kernel Hilbert spaces in probability and statistics*, Netherlands: Springer.
- Bhattacharya, A., Pati, D., and Dunson, D.B. (2011), “Adaptive dimension reduction with a Gaussian process prior,” *Arxiv preprint arXiv:1111.1044*.
- Choi, T. and Schervish, M.J. (2007), “On posterior consistency in nonparametric regression problems,” *Journal of Multivariate Analysis*, 98, 1969–1987.
- Constantine, W. and Percival, D. (2010), *wmtsa: Insightful Wavelet Methods for Time Series Analysis*, <http://CRAN.R-project.org/package=wmtsa>.
- Coombes, K., Koomen, J., Baggerly, K., Morris, J., and Kobayashi, R. (2005a), “Understanding the characteristics of mass spectrometry data through the use of simulation,” *Cancer Informatics*, 1, 41.
- Coombes, K.R., Tsavachidis, S., Morris, J.S., Baggerly, K.A., Hung, M.C., and Kuerer, H.M. (2005b), “Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform,” *Proteomics*, 5, 4107–4117.
- Cottrell, J. and London, U. (1999), “Probability-based protein identification by searching sequence databases using mass spectrometry data,” *Electrophoresis*, 20, 3551–3567.
- Crainiceanu, C.M., Ruppert, D., Carroll, R.J., Joshi, A., and Goodner, B. (2007), “Spatially adaptive Bayesian penalized splines with heteroscedastic errors,” *Journal of Computational and Graphical Statistics*, 16, 265–288.
- Denison, D.G.T., Mallick, B.K., and Smith, A.F.M. (1998), “Automatic Bayesian curve fitting,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 333–350.
- Dimatteo, I., Genovese, C.R., and Kass, R.E. (2001), “Bayesian curve-fitting with free-knot splines,” *Biometrika*, 88, 1055.
- Domon, B. and Aebersold, R. (2006), “Mass spectrometry and protein analysis,” *Science*, 312, 212.

- Donoho, D.L. and Johnstone, I.M. (1995), “Adapting to unknown smoothness via wavelet shrinkage,” *Journal of the American Statistical Association*, 1200–1224.
- Donoho, D.L. and Johnstone, J.M. (1994), “Ideal spatial adaptation by wavelet shrinkage,” *Biometrika*, 81, 425–455.
- Durbin, J. and Koopman, S. (2002), “A simple and efficient simulation smoother for state space time series analysis,” *Biometrika*, 89, 603.
- Durbin, J. and Koopman, S.J. (2001), *Time series analysis by state space methods*, vol. 24, Oxford: Oxford University Press.
- Fan, J. and Gijbels, I. (1995), “Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 371–394.
- Friedman, J.H. (1991), “Multivariate adaptive regression splines,” *The Annals of Statistics*, 1–67.
- Friedman, J.H. and Silverman, B.W. (1989), “Flexible parsimonious smoothing and additive modeling,” *Technometrics*, 3–21.
- George, E.I. and McCulloch, R.E. (1993), “Variable selection via Gibbs sampling,” *Journal of the American Statistical Association*, 881–889.
- Ghosal, S. and Roy, A. (2006), “Posterior consistency of Gaussian process prior for nonparametric binary regression,” *The Annals of Statistics*, 34, 2413–2429.
- Green, P.J. (1995), “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination,” *Biometrika*, 82, 711.
- Heckman, N.E. and Ramsay, J.O. (2000), “Penalized regression with model-based penalties,” *Canadian Journal of Statistics*, 28, 241–258.
- Kloeden, P.E. and Platen, E. (1992), *Numerical Solution of Stochastic Differential Equations*, New York: Springer Verlag.
- Lawrence, N.D., Seeger, M., and Herbrich, R. (2002), “Fast sparse Gaussian process methods: The informative vector machine,” *Advances in neural information processing systems*, 15, 609–616.

- Luo, Z. and Wahba, G. (1997), “Hybrid adaptive splines,” *Journal of the American Statistical Association*, 107–116.
- Morris, J., Coombes, K., Koomen, J., Baggerly, K., and Kobayashi, R. (2005), “Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum,” *Bioinformatics*, 21, 1764–1775.
- Morris, J.S., Brown, P.J. and Herrick, R.C., Baggerly, K.A., and Coombes, K.R. (2008), “Bayesian Analysis of Mass Spectrometry Proteomic Data Using Wavelet-Based Functional Mixed Models,” *Biometrics*, 64, 479–489.
- Neal, R. (1998), “Regression and classification using gaussian process priors,” *Bayesian Statistics*, 6, 475–501.
- Pintore, A., Speckman, P., and Holmes, C.C. (2006), “Spatially adaptive smoothing splines,” *Biometrika*, 93, 113.
- Quinonero-Candela, J. and Rasmussen, C.E. (2005), “A unifying view of sparse approximate Gaussian process regression,” *The Journal of Machine Learning Research*, 6, 1939–1959.
- R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org>.
- Rasmussen, C.E. and Williams, C.K.I. (2006), *Gaussian processes for machine learning*, Boston: MIT Press.
- Ruppert, D. and Carroll, R.J. (2000), “Spatially-adaptive Penalties for Spline Fitting,” *Australian & New Zealand Journal of Statistics*, 42, 205–223.
- Savitsky, T., Vannucci, M., and Sha, N. (2011), “Variable selection for nonparametric Gaussian process priors: Models and computational strategies,” *Statistical Science*, 26, 130–149.
- Searle, S.R. (1982), *Matrix Algebra Useful for Statistics*, New York: Wiley.
- Shepp, L.A. (1966), “Radon-Nikodym derivatives of Gaussian measures,” *The Annals of Mathematical Statistics*, 37, 321–354.

- Shi, J.Q. and Choi, T. (2011), *Gaussian Process Regression Analysis for Functional Data*, London: Chapman & Hall/CRC Press.
- Smith, M. and Kohn, R. (1996), “Nonparametric regression using Bayesian variable selection,” *Journal of Econometrics*, 75, 317–343.
- Smola, A.J. and Bartlett, P. (2001), “Sparse greedy Gaussian process regression,” in *Advances in Neural Information Processing Systems 13*, Citeseer.
- Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q.T. (2004), “Sample classification from protein mass spectrometry, by peak probability contrasts,” *Bioinformatics*, 20, 3034–3044.
- Van der Vaart, A.W. and Van Zanten, J.H. (2008a), “Rates of contraction of posterior distributions based on Gaussian process priors,” *The Annals of Statistics*, 36, 1435–1463.
- (2008b), “Reproducing kernel Hilbert spaces of Gaussian priors,” *Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, 3, 200–222.
- Van der Vaart, A.W. and Wellner, J.A. (1996), *Weak convergence and empirical processes*, New York: Springer Verlag.
- Wahba, G. (1990), *Spline models for observational data*, vol. 59, Philadelphia: Society for Industrial Mathematics.
- (1995), “Discussion of a paper by Donoho et al.” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 360–361.
- West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, New York: Springer Verlag.
- Wolpert, R.L., M.A, C., and C., T. (2011), “Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels,” *The Annals of Statistics*, 39, 1916–1962.
- Wood, S.A., Jiang, W., and Tanner, M. (2002), “Bayesian mixture of splines for spatially adaptive nonparametric regression,” *Biometrika*, 89, 513.

- Wu, J.Q., Sklar, J., Wang, Y.D., and Meiring, W. (2011), *bsml: Basis Selection from Multiple Libraries*, <http://CRAN.R-project.org/package=bsml>.
- Zhou, S. and Shen, X. (2001), “Spatially adaptive regression splines and accurate knot selection schemes,” *Journal of the American Statistical Association*, 96, 247–259.
- Zhu, B., Song, P.X.K., and Taylor, J.M.G. (2011), “Stochastic Functional Data Analysis: A Diffusion Model-Based Approach,” *Biometrics*. *In press*.
- Zou, F., Huang, H., Lee, S., and Hoeschele, I. (2010), “Nonparametric Bayesian Variable Selection With Applications to Multiple Quantitative Trait Loci Mapping With Epistasis and Gene–Environment Interaction,” *Genetics*, 186, 385.

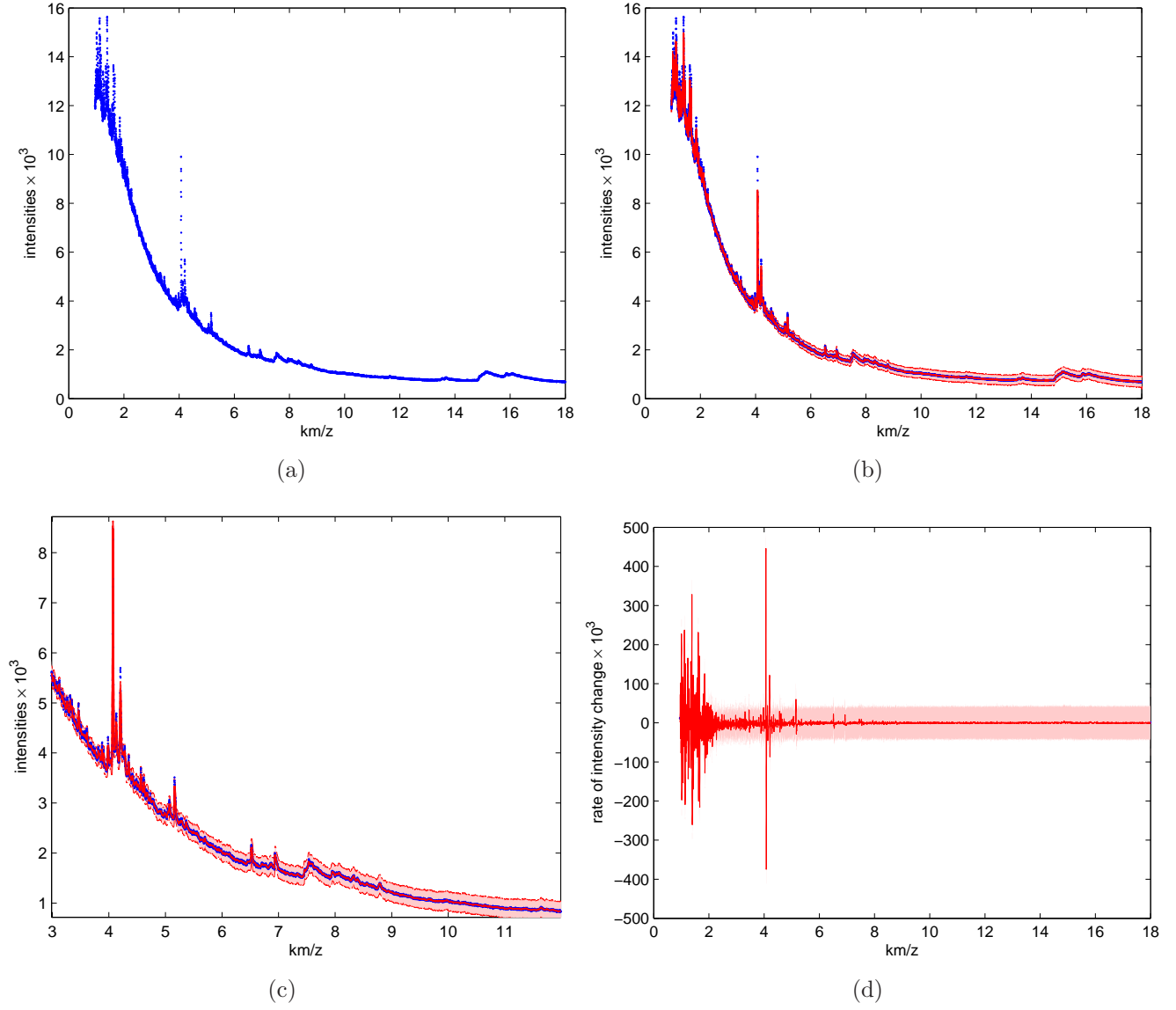


Figure 1: (a) Plot of protein mass spectrometry data: observed intensities versus mass to charge ratio m/z ; (b) Posterior mean (—) and 95% credible interval of U (red shades); (c) Posterior mean and 95% credible interval of U for a local region; (d) Posterior mean and 95% credible interval of rate of intensity changes DU .

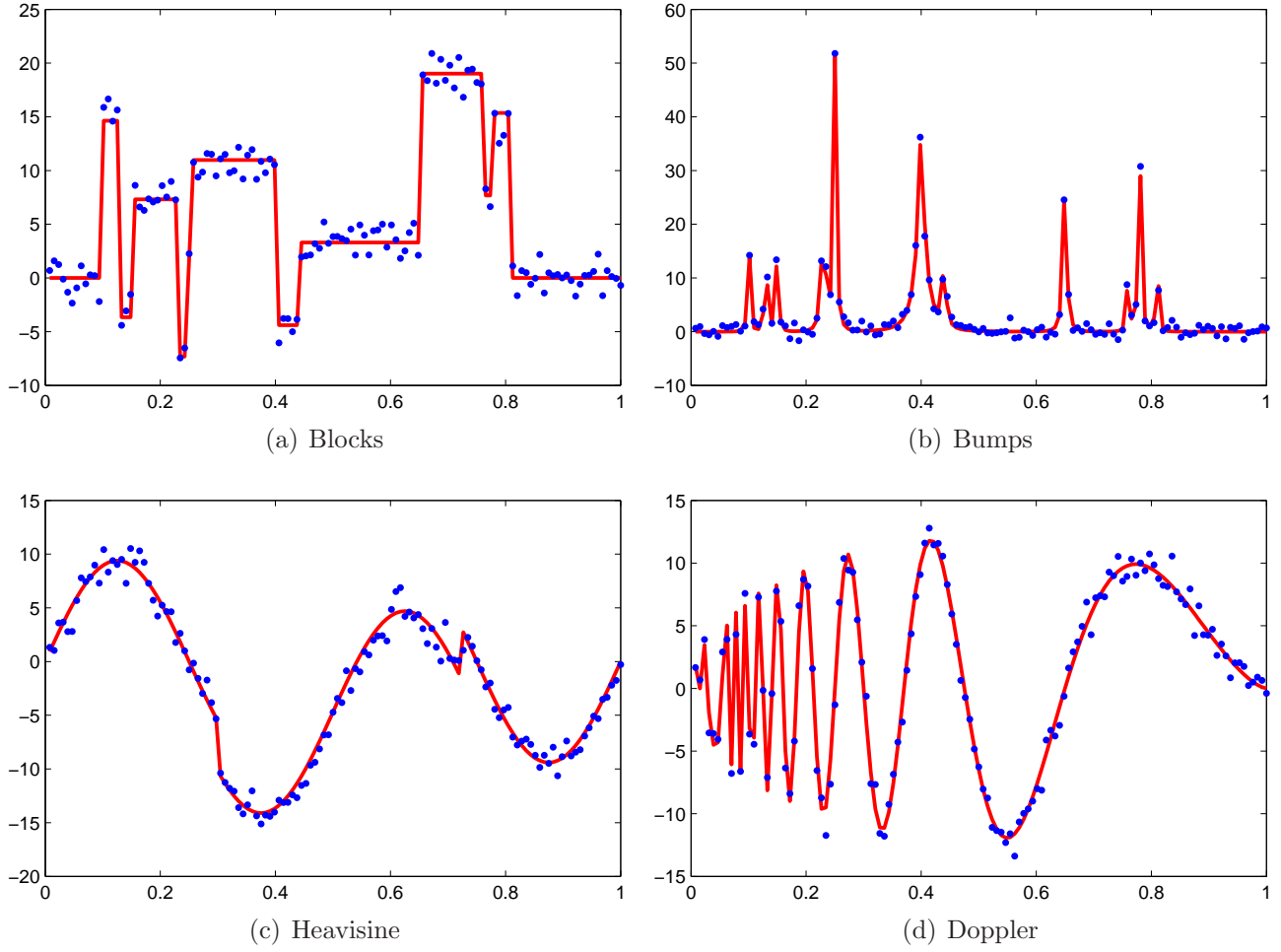


Figure 2: Four locally-varying smoothness functions: true function (—) and 128 observations (●).

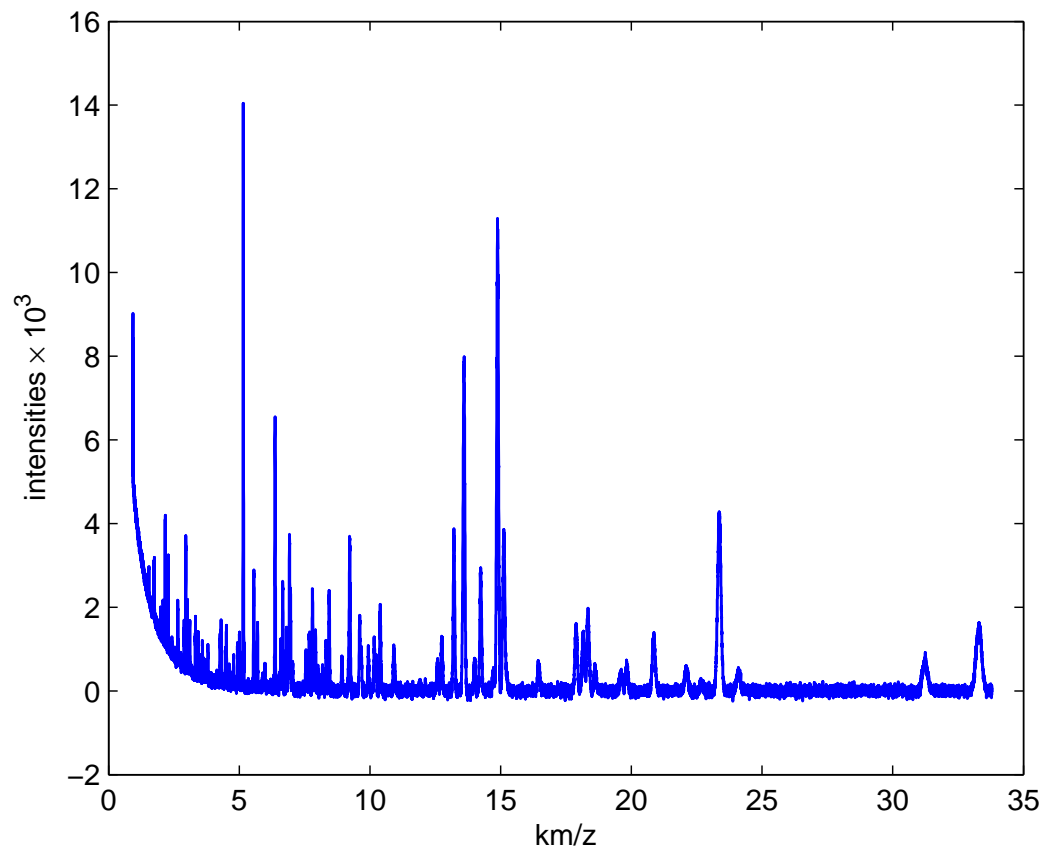


Figure 3: The plot of one simulated mass spectrometry data ($J=20,695$).

Table 1: Average MSE and the interquartile range of MSE (in parentheses) for Bayesian nonparametric regression with nGP prior (BNR-nGP), smoothing spline (SS), wavelet method with the soft minimax threshold (Wavelet1), wavelet method with the soft Stein’s unbiased estimate of risk for threshold choice (Wavelet2) and Hybrid adaptive spline (HAS).

Example	BNR-nGP	SS	Wavelet1	Wavelet2	HAS
Blocks	0.950(0.166)	3.018(0.248)	2.750(0.748)	1.237(0.341)	0.539(0.113)
Bumps	1.014(0.185)	26.185(0.787)	3.433(0.938)	1.195(0.282)	0.904(0.258)
Heavisine	0.320(0.058)	0.337(0.087)	0.702(0.230)	1.620(0.460)	0.818(0.122)
Doppler	0.989(0.183)	3.403(0.361)	1.517(0.402)	0.695(0.179)	3.700(0.534)
MS Region 1($\times 10^{-3}$)	1.498(0.266)	2.293(0.513)	2.367(0.616)	6.048(3.441)	72.565(39.596)
MS Region 2($\times 10^{-3}$)	0.840(0.375)	0.798(0.490)	0.948(0.587)	1.885(0.493)	7.958(5.559)